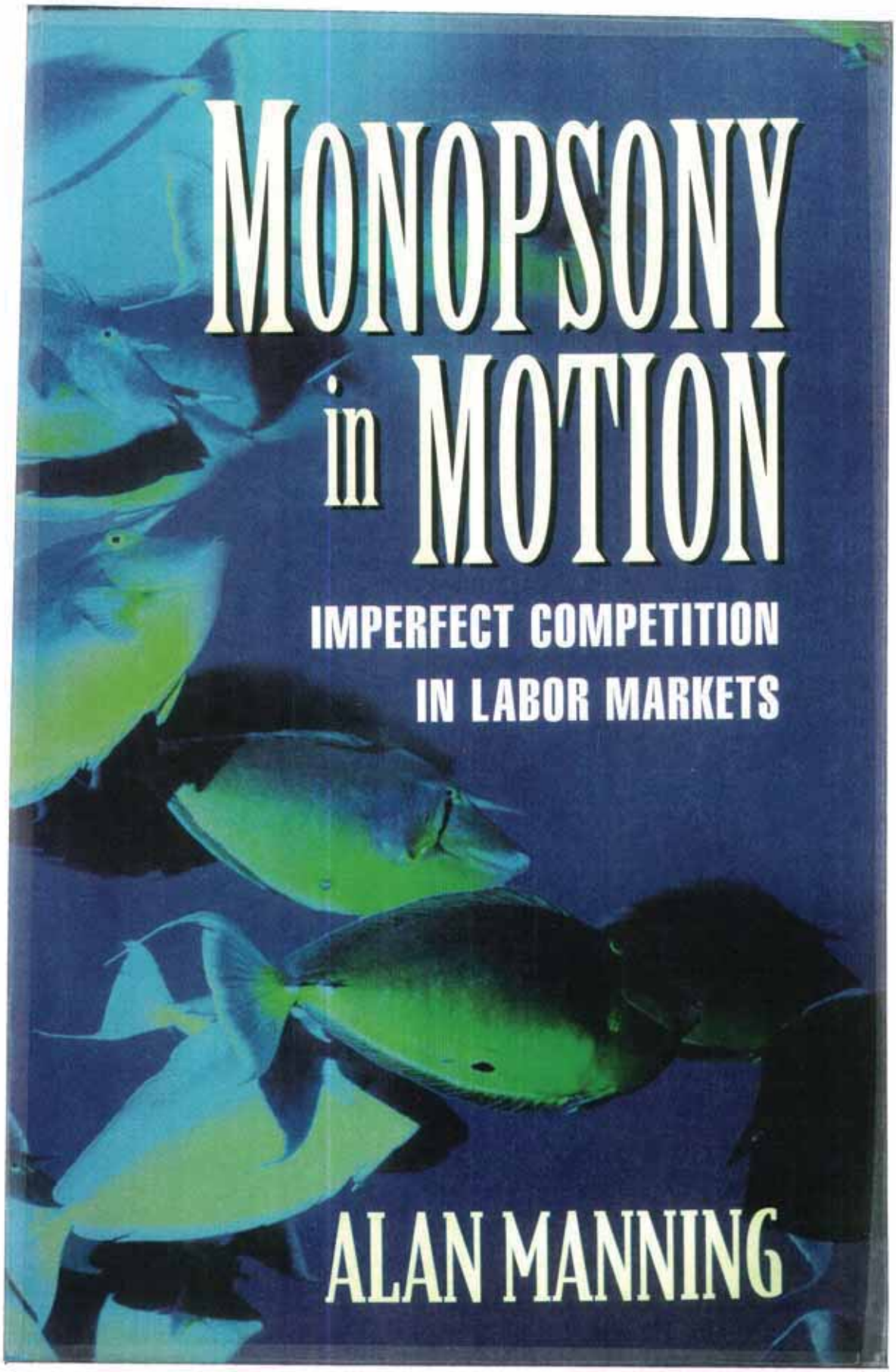
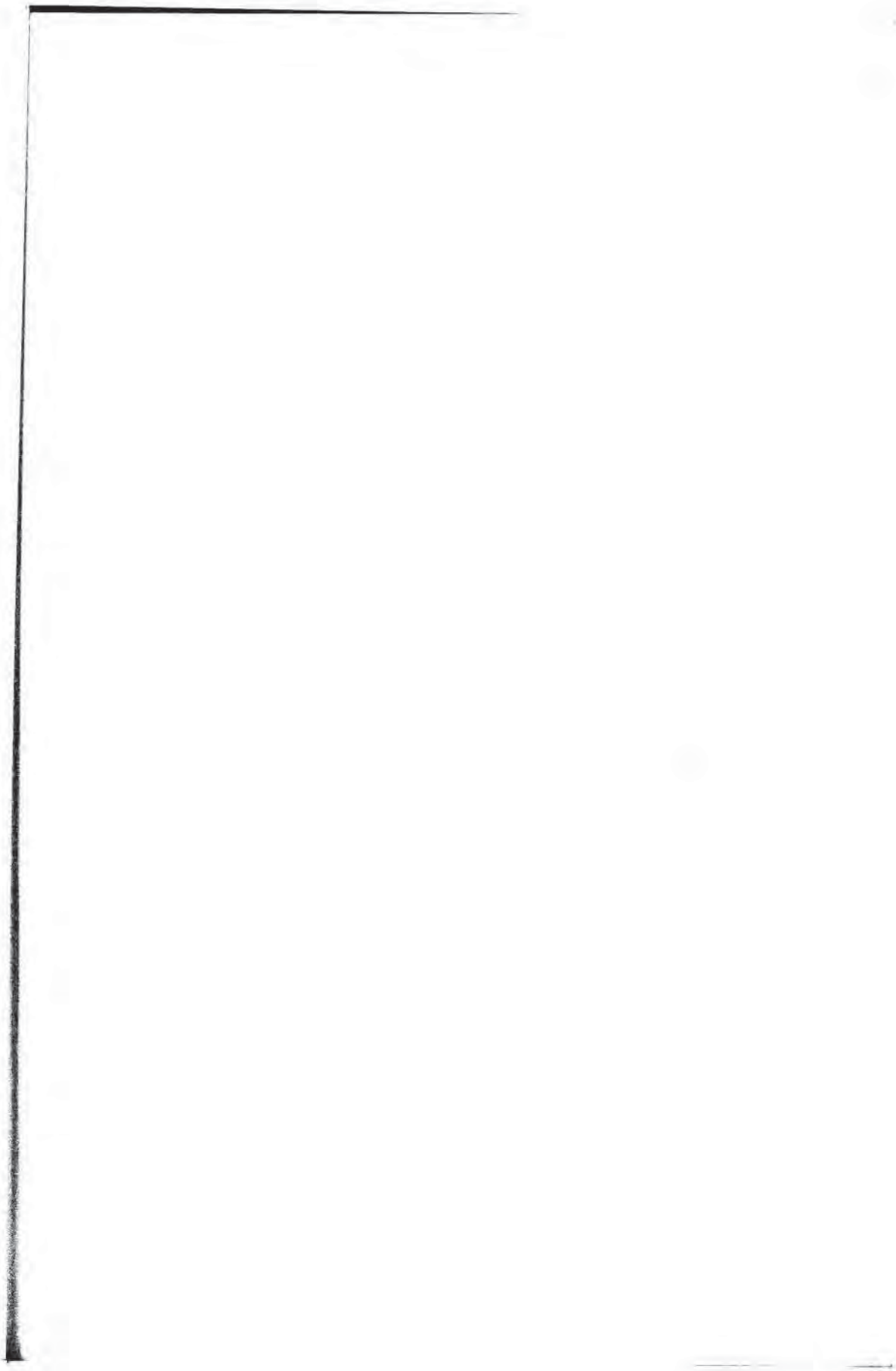
Joint Ex. 1

(JCCX33 - Alan Manning, Monopsony in Motion (Princeton University Press, 2003))

PART 1


MONOPSONY
in MOTION
IMPERFECT COMPETITION
IN LABOR MARKETS
ALAN MANNING

MONOPSONY IN MOTION



MONOPSONY IN MOTION

IMPERFECT COMPETITION IN LABOR MARKETS

Alan Manning

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
3 Market Place, Woodstock, Oxfordshire OX20 1SY

Library of Congress Cataloging-in-Publication Data applied for

ISBN 0-691-11312-2 (alk. paper)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

This book has been composed in Sabon

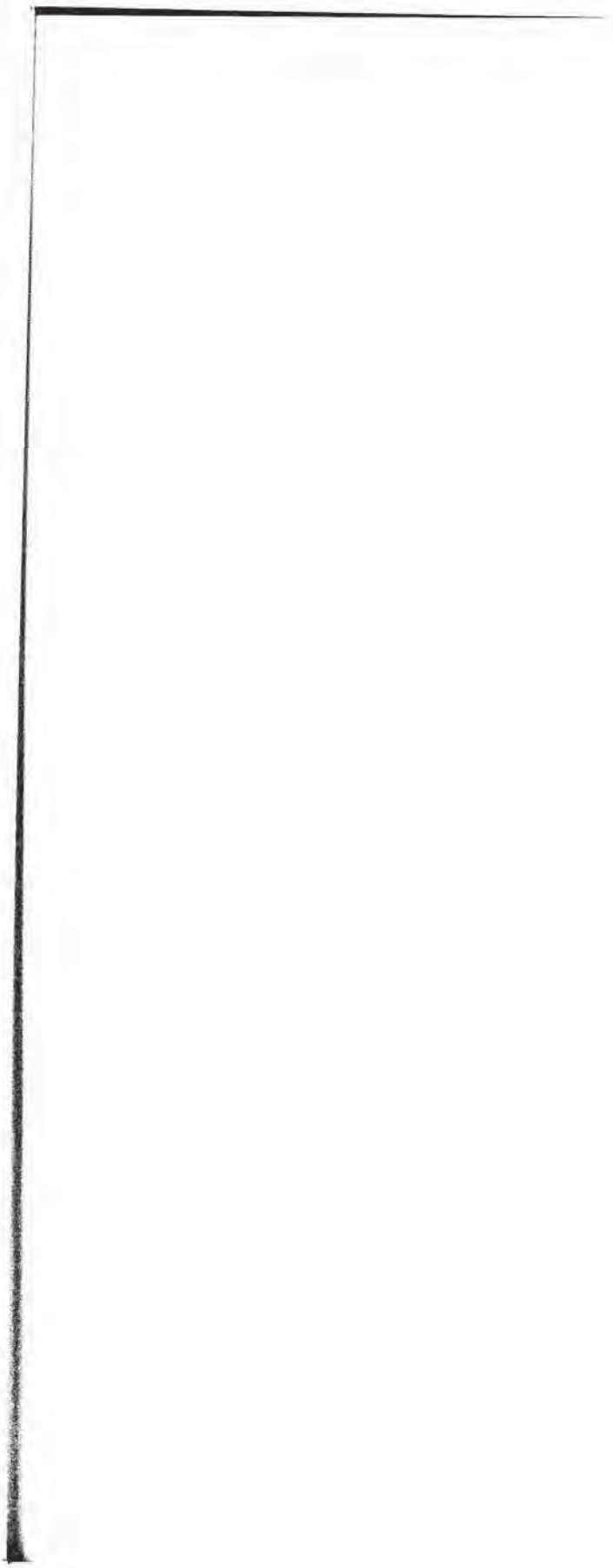
Princeton University Press books are printed on acid-free paper, and meet the guidelines for the permanence and durability of the Committee on Production Guidelines for Book Longevity of the Council on Library Resources

www.pupress.princeton.edu

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my family



Contents

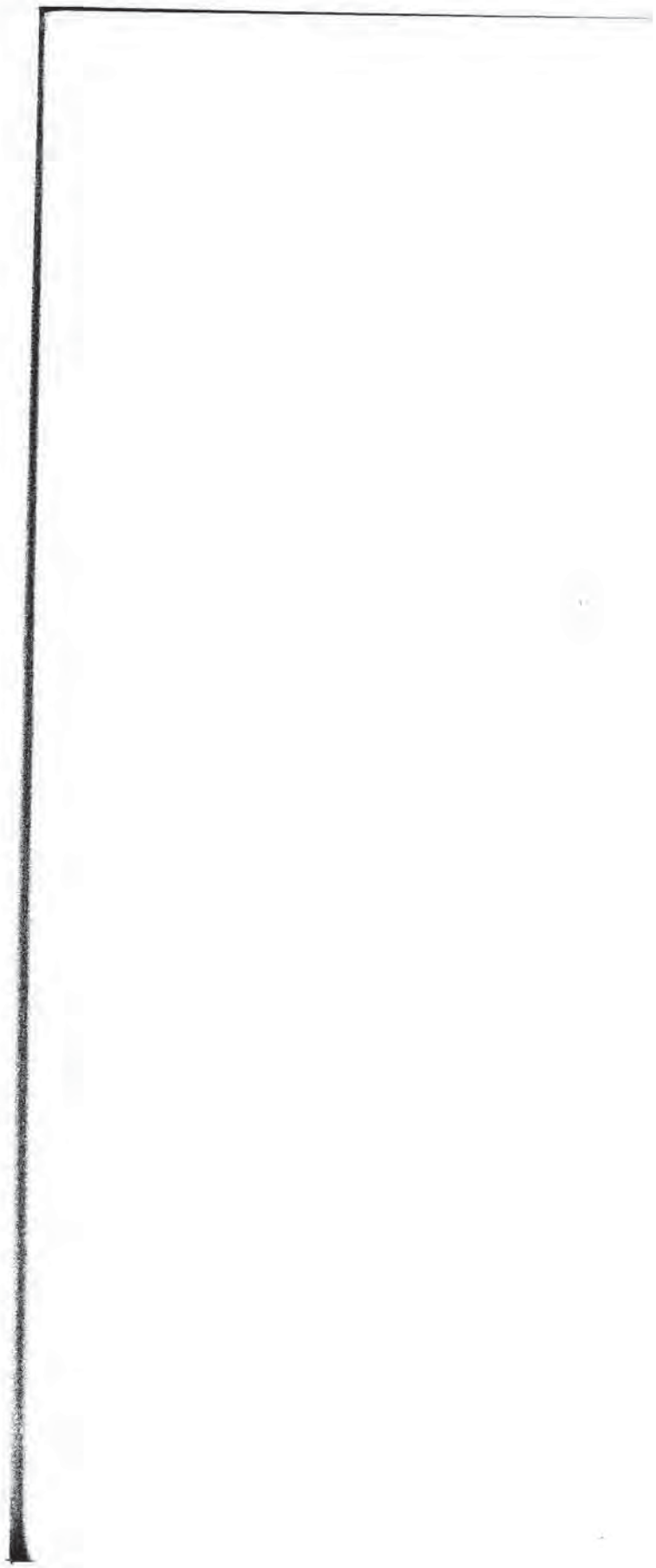
Preface

I used to be like most other labor economists and think that the textbook model of monopsony was little more than a curiosum, perhaps relevant in a few times and places but not of much use for thinking about most labor markets. That view began to change after seeing a presentation of Burdett and Mortensen (1998) at the Centre for Economic Performance, LSE, in 1990. That made me realize that the assumption that employers have no market power over their workers is equivalent to the wildly implausible conclusion that a wage cut of a cent causes all existing workers to leave the firm immediately. I thought that the idea that frictions in the labor market give employers some power over their workers deserved more thought than was usually given to it. Those ideas then evolved into the belief that a perspective on labor markets based on the view that "monopsony" is important led to a much better understanding of a very wide range of labor market phenomena. And, eventually, this book is the result of the attempt to substantiate that belief.

Eventually, because this book took too much time to write. Along the way, I have been generously supported in time and money by many people. In particular David Card who arranged for me to spend time at the Industrial Relations Section, Princeton in 1994–1995 and the Center for Labor Economics, University of California Berkeley in 1998–1999, and who gave me so many valuable comments and so much support. When he asked me how it was going every time we met, my embarrassment at the lack of progress spurred me on. And, when he no longer enquired about the book, the guilty feeling that he had given up on me caused even greater embarrassment. He arranged a one-day conference on the bulk of the first draft in Berkeley in July 1999 that provided incredibly helpful comments: I would like to thank the other participants, Paul Beaudry, Marianne Bertrand, John diNardo (who also wrote, arranged and performed the song "Monopsony in Motion" that, in this multi-media age, goes with the book and can be found on my website), Steve Machin, Dale Mortensen, Sendhil Mullanaithan, Michael Ransom, and Craig Riddell.

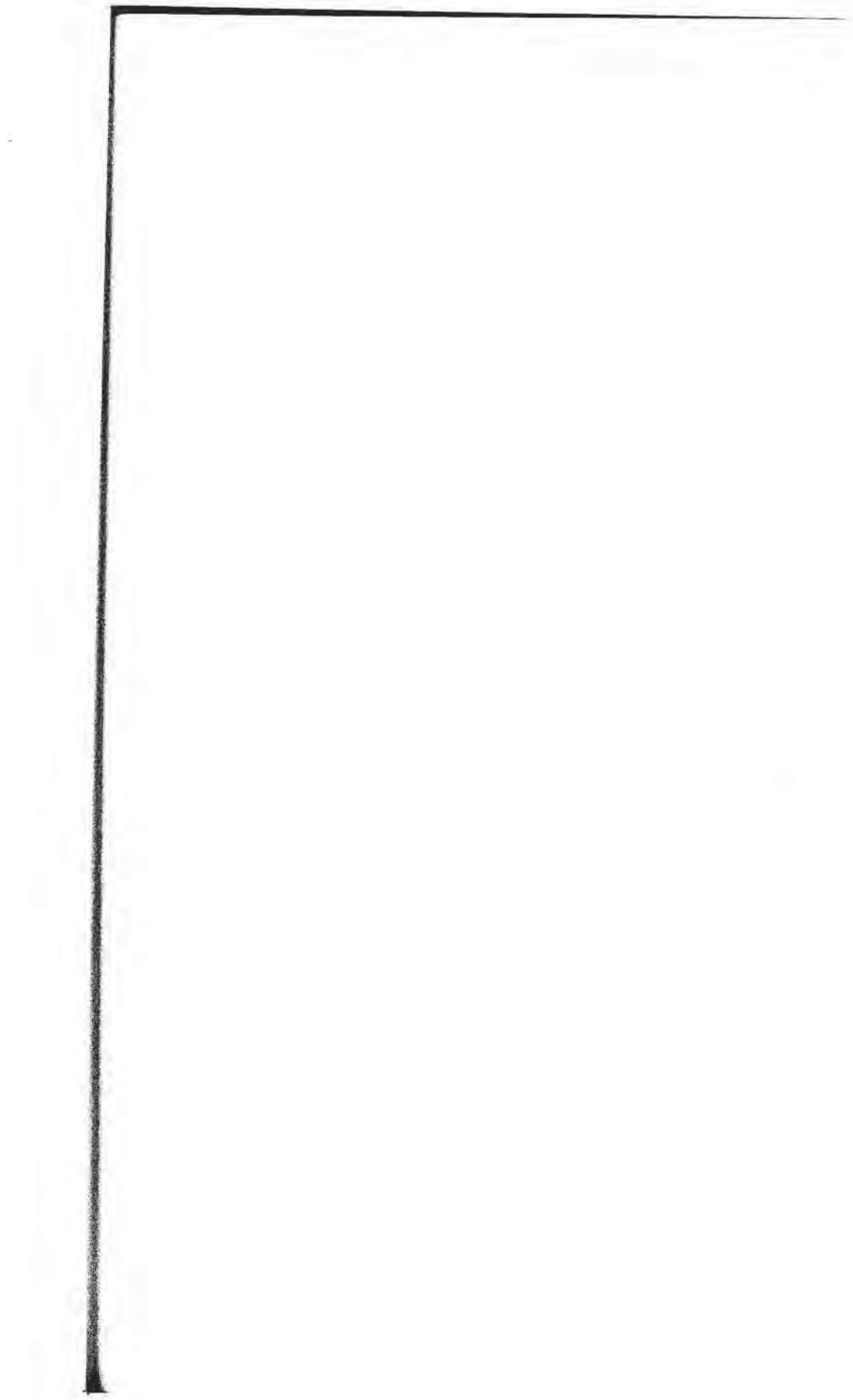
Richard Layard of the Centre for Economic Performance at the LSE also generously provided a teaching buy-out that gave me a much-needed breathing space to work on the book.

Others who provided comments on the book are (in alphabetical order) Damon Clark, Richard Disney, Juan Dolado, Maarten Goos, Maia Guell, Marco Manacorda, Karl Ove Moene, Barbara Petrongolo, Jumana Saleheen, Steve Pischke, and Coen Teulings. Pat Nutt helped me avoid the tedious task of assembling the bibliography.



Part One

BASICS



1

Introduction

WHAT happens if an employer cuts the wage it pays its workers by one cent? Much of labor economics is built on the assumption that all existing workers immediately leave the firm as that is the implication of the assumption of perfect competition in the labor market. In such a situation an employer faces a market wage for each type of labor determined by forces beyond its control at which any number of these workers can be hired but any attempt to pay a lower wage will result in the complete inability to hire any of them at all. The labor supply curve facing the firm is infinitely elastic.

In contrast, this book is based on two assumptions about the labor market. They can be stated very simply:

- there are important frictions in the labor market;
- employers set wages.

The implications of these assumptions can also be stated simply. The existence of frictions means that there are generally rents to jobs: if an employer and worker are forcibly separated one or, more commonly, both of the parties would be made worse off. This gives employers some market power over their workers as a small wage cut will no longer induce them to leave the firm. The assumption that employers set wages then tells us that employers exercise this market power. But, with these two assumptions, it is monopsony, not perfect competition, that is the best simple model to describe the decision problem facing an individual employer. Not monopsony in the sense of there being a single buyer of labor, but monopsony in the sense of the supply of labor to an individual firm not being infinitely elastic. The actions of other employers (notably their choice of wages) in the market will affect the supply of labor to an individual firm so, if one wants to model the market as a whole, models of oligopsony or monopsonistic competition are what is needed.[1] The usefulness of the monopsonistic approach rests on the two assumptions so they need some justification.

[1] The Oxford English Dictionary credits the word monopsony to Joan Robinson (1933) though she credits it to B. L. Hallward, a classical scholar at Cambridge, who though born in 1901 is still alive at the time of writing. The suffix is derived from OPSONEN which means "to make your purchases often of dried fish" and which is found in Aristophanes, the Wasps (twice), Plutarch and the New Testament. The natural ONEOMAI ("I buy") was rejected as it does not sound good with the MONO prefix (personal communication to David Card). The invention of the word oligopsony is credited to Walker (1943) who introduced it with the curious phrase "it is surely only a matter of time before market situation number 23 is christened oligopsony," the time referred to being the time necessary for him to finish writing the sentence.

That important frictions exist in the labor market seems undeniable: people go to the pub to celebrate when they get a job rather than greeting the news with the shrug of the shoulders that we might expect if labor markets were frictionless. And people go to the pub to drown their sorrows when they lose their job rather than picking up another one straight away. The importance of frictions has been recognized since at least the work of Stigler (1961, 1962).

What are the sources of these frictions in labor markets? In the *Economics of Imperfect Competition*, Joan Robinson argued that:

> there may be a certain number of workers in the immediate neighbourhood and to attract those from further afield it may be necessary to pay a wage equal to what they can earn near home plus their fares to and fro; or there may be workers attached to the firm by preference or custom and to attract others it may be necessary to pay a higher wage. Or ignorance may prevent workers from moving from one to another in response to differences in the wages offered by the different firms.
>
> (Robinson 1933: 296)

It is ignorance, heterogeneous preferences, and mobility costs that are the most plausible sources of frictions in the labor market. The consequence of these frictions is that employers who cut wages do not immediately lose all their workers. They may find that their workers quit at a faster rate than before or that recruitment is more difficult, but the extreme predictions of the competitive model do not hold. The labor supply curve facing the firm is, as a result, not infinitely elastic. The existence of frictions gives employers potential market power over their workers. The assumption that firms set wages means that they actually exercise this power. Let us now consider this assumption in more detail.

Given the existence of rents caused by frictions one needs to specify how they are divided between employer and worker. The existence of the rents makes the relationship between workers and employer one of bilateral monopoly (in part) so that we need a theory of how the rents are divided. The development of such a theory is an old problem in economics in general, and labor economics in particular, going back to the discussion of Edgeworth (1932) who argued that the terms of exchange in bilateral monopoly were indeterminate. This indeterminacy has never been resolved.[2]

[2] For example, in recent years, there has been considerable interest in models of bargaining in bilateral monopoly following on from the work of Rubinstein (1982) who, for a particular specification of the negotiation process between the two parties, showed that there was a unique equilibrium, that is, a determinate outcome. But this literature does not really solve the indeterminacy problem, it just pushes it back one more stage for the rules of the negotiation process generally determine the division of the rents and these rules are essentially arbitrary. So the indeterminacy problem re-emerges in the indeterminacy of the rules of the game.

Given this problem at the heart of economics, which this book is going to make no attempt to solve, there seems little alternative but to grasp the nettle and make some assumption about the way in which the rents are divided. One should choose an assumption that is a reasonable approximation to reality. This is made difficult by the fact that there is no universally right assumption for how rents are shared in the labor market: there are different mechanisms in different labor markets, perhaps even co-existing in the same labor market. In spite of this, we focus on one mechanism for most of this book.

In this book, it is assumed that employers set wages.[3] This is a more appropriate assumption in some labor markets than others. For example, it would not seem to be appropriate when workers are organized into a union (the consequences of this are discussed in chapter 12), or for senior management who often seem to have considerable ability to set their own wages, or for the self-employed, or (most importantly of course) for academic labor economists. But, for the average worker in a non-union setting, this does seem to be the appropriate assumption. Open the pages of a newspaper and one sees firms advertising jobs at given wages. One also sees advertisements saying "salary negotiable" though typically only for higher level jobs and the extent to which they are actually negotiable is often rather limited. But it is very rare to see advertisements placed by workers setting down the wage at which they are prepared to work.

This view that the relationship between the employer and worker is one-sided has a long tradition. In the *Wealth of Nations*, Adam Smith (1976: 84) wrote that "in the long run the workman may be as necessary to his master as his master is to him; but the necessity is not so immediate." And Alfred Marshall in his *Principles of Economics* (1920: 471) wrote that "labour is often sold under special disadvantages arising from the closely connected group of facts that labour power is 'perishable', that the sellers of it are commonly poor and have no reserve fund, and that they cannot easily withhold it from the market." To these arguments that a worker is typically more desperate for work than an employer is desperate for that particular worker, Sidney and Beatrice Webb (1897: 657–58) added the argument that

> the manual worker is, from his position and training, far less skilled than the employer ... in the art of bargaining itself. This art forms a large part of the daily life of the entrepreneur, whilst the foreman is specially selected for his skill in engaging and superintending workmen. The manual worker, on the

[3] Section 1.3 compares this assumption about wage setting with a prominent alternative, the ex post bargaining used in much of the matching literature (for a recent survey, see Mortensen and Pissarides 1999).

> contrary, has the smallest experience of, and practically no training in, what is essentially one of the arts of the capitalist employer. He never engages in any but one sort of bargaining, and that only on occasions which may be infrequent, and which in any case make up only a tiny fraction of his life.

The view that the relationship between employer and worker is not one of equals was the origin of pro-labor legislation in many if not all countries. Section 1 of the US National Labor Relations Act of 1935 says "the inequality of bargaining power between employees who do not possess freedom of association or actual liberty of contract, and employers who are organized in the corporate and other forms of ownership association substantially burdens and affects the flow of commerce." Our assumption that employers set wages is in this tradition.

The claim that labor markets are, in the absence of outside intervention, pervasively monopsonistic probably comes as something of a surprise to readers of labor economics textbooks. Table 1.1 documents the number of pages devoted to a discussion of monopsony and the total length in a selection of popular textbooks. As can be seen, monopsony does not figure prominently and, where it is mentioned, the discussion is generally not favorable: the final column of table 1.1 contains a selection of quotes, some of which capture the idea that frictions give employers some market power but most of which do not.[4] There is a noticeable trend in the most recent textbooks towards less hostile views[5] and a recognition that it is the existence of labor market frictions that is the main argument for the relevance of monopsony. But, while the overall perspective on the plausibility of monopsony may be changing, the range of labor market issues that contain some discussion of the implications of monopsony remains very limited. The first two volumes of the *Handbook of Labor Economics* (Ashenfelter and Layard, 1986) contain only two references to monopsony out of a total of 1268 pages, one in the chapter on dynamic labor demand by Nickell and the other in the chapter on discrimination by Cain. The three subsequent volumes published in 1999 (Ashenfelter and Card, 1999) contain three references in 2362 pages, in the chapters on labor market institutions, minimum wages and matching.

[4] My personal favorite is taken from the first edition of Fleisher and Kniesner (1980, pp. 203–4) "we feel confident that monopsony is not a widespread phenomenon today. The primary reason is that fame and financial awards await the researcher who can demonstrate empirically that a significant number of workers are victims of monopoly power of employers. As yet, no one has claimed these prizes."

[5] A trend that can be confirmed by a fixed effect estimator, comparing the discussion of monopsony in different editions of the most popular textbooks.

TABLE 1.1
Monopsony in Labor Textbooks

Author	*Pages on Monopsony*	*Total Pages*	*Selected Quotation*
Borjas (2000)	7	470	"upward-sloping supply curves for particular firms can arise even when there are many firms competing for the same type of labour" (p. 191)
Ehrenberg and Smith (2000)	14	651	"while examples of a single buyer of labour services may be difficult to cite, the monopsony model still offers useful insights if the labour supply curves are upward-sloping for some other reason. Recently, economists have begun to explore a variety of labour market conditions that would yield upward-sloping labour supply curves to individual firms even when there are many employers competing for workers in the same labour market" (p. 71)
Filer, Hamermesh, and Rees (1996)	8	654	"it does not seem plausible given the vast number of firms employing teenagers in the US" (p. 174) "while the cost of commuting long distances leaves some residual monopsony power to isolated employers, this power is much less than when commuting was more difficult" (p. 189)
Sapsford and Tzannatos (1993)	15	420	
Elliott (1991)	6	536	"appealing as such an outcome is to the advocates of minimum wage legislation it has to be said that this theoretical possibility is seldom encountered in practice" (p. 306).
Kaufman (1991)	12	778	"the pure form of monopsony (the one-company town) is relatively rare, although conditions of oligopsony and monopsonistic competition may have a wider applicability" (pp. 422–23)
Reynolds, Masters, and Moser (1991)	2	610	"there is little evidence to suggest that monopsony is important to our economy. Most firms are located in urban areas where there are many firms in the labour market and relatively little collusion among employers" (p. 135)

TABLE 1.1 (*continued*)

Author	*Pages on Monopsony*	*Total Pages*	*Selected Quotation*
Fallon and Verry (1988)	3	311	"imperfect information may ... convey some monopsony power to the individual firm" (p. 103)
Gunderson and Riddell (1988)	19	600	"to a certain extent most firms may have an element of monopsony power in the short-run, in the sense that they could lower their wages somewhat without losing all their workforce. However, it is unlikely that they would exercise this power in the long run because it would lead to costly problems of recruitment, turnover and morale" (pp. 213–14) "Improved communications, labour market information, and labour mobility make the isolated labour market syndrome, necessary for monopsony, unlikely at least for large numbers of workers" (p. 224)
Hoffman (1986)	7	354	"A monopsonist is a firm that faces an upward-sloping supply curve for labor of a given quality. A university hiring economics instructors is most definitely *not* a monopsonist, because the relevant labor market is national and thus the number of other demanders is quite large" (p. 49)
McConnell and Brue (1986)	9	607	"monopsony outcomes are not widespread in the US economy" (p. 150)
Marshall, Briggs, and King (1984)	4	657	
Fleisher and Kniesner (1984)	16	536	"monopsony does not appear to be a widespread phenomenon in the United States, but rather specific to a few industries" (p. 219)
Hunter and Mulvey (1981)	4	403	
Fearn (1981)	8	272	"many modern American labor economists assume generally that labor markets are competitive. The presumption that labor markets are monopsonies, however, remains in the public consciousness, particularly in union circles and in the legislatures. The situation may represent a classic 'cultural lag'" (p. 117)

TABLE 1.1 (*continued*)

Author	*Pages on Monopsony*	*Total Pages*	*Selected Quotation*
Bloom and Northrup (1981)	4	836	
Kreps, Martin, Perlman, and Somers (1980)	9	477	"instances of monopsony are not that frequent as to make the chances that administered wages will not reduce employment" (p. 110)
Addison and Siebert (1979)	8	500	"we should qualify our discussion of monopsony by observing that imperfect worker information as to alternative wages will confer on each firm a margin of monopsony power. Thus, each firm will possess a degree of dynamic monopsony power arising from the imperfect information of its employees and can therefore administer wages" (p. 169)
Freeman (1979)	0	196	
Bellante and Jackson (1979)	4	351	"many economists argue that monopsony power by firms is likely to be greatly exaggerated given the occupational, industrial and geographical mobility that characterizes American labor markets" (p. 196)
Cartter and Marshall (1972)	11	570	
Lester (1964)	2	608	"the manipulation of wages by the purchase of labor according to monopsonistic calculations seems to be misguided academic speculation" (p. 281)
Phelps Brown (1962)	1	274	"the rate needed to attract labour in the first place is higher than that needed to retain it once it has settled in. Much of a firm's labour force is likely, for this reason, to be captive; the firm is a monopsonist in the short-run' (p. 137)

These statistics might be thought to be a little unfair as many of these textbooks interpret monopsony literally as being a situation of a single employer of labor rather than the interpretation of an upward-sloping supply curve of labor that is used here. But, mentions of oligopsony are even fewer than mentions of monopsony, and the

general impression given by most textbooks is that employers have negligible market power over their workers or that this is, at best, a trivial side issue.

This situation contrasts strongly with the situation in another part of economics, industrial organization, where the standard assumption is that all firms have some product market power, although some are thought to have more market power than others. As a result, the bulk of the *Handbook of Industrial Organization* is about imperfect competition in product markets and virtually every chapter has some reference to monopoly or oligopoly. This contrast between labor economics and industrial organization is odd given that one might think frictions are more important in the labor market as it is more costly to change one's job than one's supermarket.[6] The premise of this book is that labor economics should adopt a similar attitude to that in industrial organization and start analysis from the position that all employers have some labor market power.

This book discusses most if not all of the issues in labor economics from the starting point that the labor market is monopsonistic. Given the evidence cited above on the paucity of references to monopsony in textbooks, one might expect a radical reworking of labor economics. Such an expectation will, more often than not, lead to disappointment. Often, we will be able to draw heavily on existing work and simply look at issues from a different angle. Many explanations of labor market phenomena implicitly assume that the labor market is monopsonistic without articulating that fact. Perhaps the best example of this is search theory, an approach used to analyze a wide range of issues. The early developments, following Stigler (1962), were one-sided, treating the distribution of wage offers in the market as exogenous. Stigler (1962) provides a careful and interesting discussion of why the "law of one wage" is likely to fail in the labor market but does not consider the process of wage setting from the perspective of employers. But, when the process of wage determination was considered, the early models often seemed to collapse, and were incapable of explaining the existence of a non-degenerate wage distribution, a point made forcefully by Diamond (1971) and Rothschild (1973). All of the models then developed to explain the existence of equilibrium wage dispersion (e.g., Butters 1977) essentially assume that firms have some market power. It would be an exaggeration to say

[6] For example, individuals in the British Household Panel Survey commonly report employment-related events as major life events but none report that one of the most important things that happened to them in the past year is that they stopped shopping at Sainsburys and started going to Tesco, two of the biggest British supermarkets.

that all coherent models of frictions imply firms have some market power but it is close to the truth.[7]

1.1 The Advantages of a Monopsonistic Perspective

The main advantage of the monopsonistic approach is that the way one thinks about labor markets is more "natural" and less forced. Currently, labor economics consists of the competitive model with bits bolted onto it when necessary to explain away anomalies. The result is often not a pretty sight. A good example is the analysis of the returns to specific human capital. If one is a strict believer in perfectly competitive markets, one should believe that workers get no return from firm-specific human capital: as Becker (1993: 41–42) puts it "one might plausibly argue that the wage paid by firms would be independent of training." But, Becker goes on to argue that employers need to give workers some share to "deter quits," an idea formalized by Hashimoto (1981) which is the standard reference for this conclusion. But (and this is discussed in more detail in chapter 5), Hashimoto simply assumes that the supply of labor to the firm is not perfectly elastic, that is, he assumes the labor market is monopsonistic, a rather helpless fudge that has sown only confusion ever since.

Assuming labor markets are monopsonistic also brings the thinking of labor economists in line with the way in which agents perceive the workings of labor markets. Workers do not perceive labor markets as frictionless and changing, getting, or losing a job are routinely reported as major life events: for example, in the UK British Household Panel Survey (BHPS), job-related events are the most common category of self-reported important life events after births, deaths, and weddings. And, employers perceive they have discretion over the wages paid. Human resource management textbooks routinely state that the choice

[7] It is instructive to consider the models of frictions that do not give employers some market power. In the "islands" model of Lucas and Prescott (1974), workers must make a decision about the island on which to work before the realization of island-specific demand shocks. There are frictions as there is no ex post mobility between islands after the realization of the shocks. But, even though there are frictions, workers get paid their marginal product as Lucas and Prescott use a "wildebeest" model of the labor market in which each island has huge herds of employers who bid the wage up to the marginal product. Somewhat similar are the models of Moen (1997), Acemoglu and Shimer (2000), and Burdett et al. (2001) where each "island" has only one firm, workers have an ex ante free choice of islands and the uncertainty about the demand shock is replaced by a matching friction in which it is hard to get employment once one is on an island. However, it is assumed that each firm commits in advance the wage it is going to pay so that the relevant labor supply curve is the perfectly elastic ex ante supply curve rather than the completely inelastic ex post one.

of the wage affects the ability of the employer to recruit and retain workers (see, e.g., Jackson and Schuler 2000, chapter 10) and the choice of a wage is a very real one.[8]

It is simple to give examples of how a monopsonistic perspective makes life more comfortable for labor economists. The existence of wage dispersion for identical workers can readily be explained as the natural outcome of a labor market in which the competitive forces are not so strong as to make it impossible for low-wage employers to remain in existence: no recourse is needed to "unobserved ability" to deny the existence of the phenomenon. When we find that workers paid (other things being equal) higher wages are less likely to be looking for another job or that they are less likely to leave their employers, this can be readily explained by the fact that these workers have been lucky enough to find themselves in one of the good jobs in their segment of the labor market. It does not have to be explained away in terms of higher-wage workers having more specific human capital (see, e.g., Neal 1998).

Similarly, the robust empirical correlation between employer characteristics and wages does not have to be explained away in terms of unobserved worker quality: it is exactly what one would expect to find in a monopsonistic labor market. When one observes that employers pay for general training for their workers, one does not have to claim that such training is really specific or that workers are paying for it indirectly. It is what one would expect in a monopsonistic labor market in which part of the returns to general training will accrue to employers.

When we find that equal pay legislation substantially raises the pay of women, and does not appear to harm their employment, this is readily explained by a monopsonistic perspective but a serious problem if one believes the labor market is perfectly competitive. Similarly, finding that the minimum wage does not harm employment prospects in some situations is no particular mystery if one believes in monopsony.

Other examples can be added and are discussed at various points in this book. But, many labor economists instinctively feel very uncomfortable with the idea that labor markets may be pervasively monopsonistic and the next section tries to allay some of these fears.

[8] Issues of labor quality muddy this as, in a competitive labor market, the choice of a wage is really the choice of quality of labor to employ on a job. But, if the competitive model of a labor market was correct, a firm that pays all its workers on a particular job the same wage (such firms are easy to find; see chapter 5) should have no variation in quality among these workers. There would be no such thing as a "most-valued" worker. However, employers are aware that there is heterogeneity in the quality of workers who are paid the same wage. So, it is probably best to think of the wage paid as affecting both the quantity and quality of workers; see Manning (1994b) for the working out of a model with this feature.

1.2 Objections to Monopsony and Oligopsony

Many labor economists find the claim that labor markets are pervasively monopsonistic inherently implausible. It is doubtful that anyone would claim literally that the labor supply curve facing a firm is, in the short run, infinitely elastic as the perfectly competitive model assumes. Almost certainly, most labor economists think of the elasticity as "high" and that the competitive model provides a tolerable approximation to reality. But, once one concedes that the competitive model is not literally true, it becomes an empirical matter just how good an approximation it is. The claim of this book is that, for many questions, the competitive model is not a tolerable approximation, and that our understanding of labor markets would be much improved by thinking in terms of a model where the labor supply curve facing the firm is not infinitely elastic.

The belief that the elasticity of the labor supply curve facing a firm is infinitely elastic is not based on any great weight of accumulated empirical evidence. The number of papers written about the elasticity of the labor supply curve at firm level can almost be counted on the fingers of one hand (see the discussion in chapter 4). Rather, it is introspection (or revelation) which is the source of the faith of many labor economists in the irrelevance of monopsony.

There are a number of sources of this faith. First, there is the belief that large employers are necessary for employers to have some market power and that the vast majority of employers are small in relation to their labor market; Bunting (1962) is the classic reference for US evidence on this. But the approach developed in this book does not require employers to be "large" in relation to their labor market. It only requires that a wage cut of a cent does not cause all workers to leave employment immediately.

Secondly, some labor economists argue that labor turnover rates are so high that workers cannot be thought of as "tied" to firms. But, the *level* of labor turnover is irrelevant: the issue is the *sensitivity* of labor turnover rates to the wage. Existing studies of this find that separations are related to the wage but that the elasticity is not enormous (again, this literature is discussed further in chapter 4).

Some other labor economists think that the supply of labor to a firm is irrelevant because they believe that the normal state of affairs is that employers are turning away workers who want a job at prevailing wages. Involuntary unemployment might be taken as one piece of evidence in this respect, low vacancy rates as another. But, we argue (in chapter 9) that the existence of monopsony and involuntary unemployment are essentially orthogonal issues. Employers have market power over their workers whenever the elasticity of the supply of workers that the employer might

consider employing is less than infinite, while involuntary unemployment exists when the supply of the workers that the employer would want is less than the supply who would like to work at the going wage.

And, we argue (in chapter 10) that low vacancy rates and durations are perfectly consistent with the existence of labor supply being a constraint on employers. As job creation is costly, firms will not create jobs they do not expect to be able to fill. Hence, one should think of vacancies as "accidents" and a low vacancy rate is perfectly consistent with employers having some monopsony power.

Thus, the faith that so many labor economists have in the irrelevance of monopsony or oligopsony is not based on hard evidence, and the throw-away arguments sometimes heard are not as compelling as generations of labor economists have been led to believe. The idea deserves to be given more serious consideration and that is the aim of this book.

In much of the previous discussion, the idea of a monopsonistic labor market has been compared to the ideal of a frictionless labor market. But, there are other labor market models which acknowledge the existence of frictions yet would not commonly be thought of as monopsony models. Perhaps the most prominent example of these models is the Diamond–Pissarides matching model (see Diamond 1982; Pissarides 1985). How these models relate to the monopsony model is the subject of the next section.

1.3 Monopsony or Matching or Both?

Another tradition in labor economics, commonly called matching models (for a recent survey, see Mortensen and Pissarides 1999), also starts from the premise that there are important frictions in labor markets. But, these models differ from monopsony models in the assumptions made about wage determination. There are two main such differences (for an explicit formal comparison of the two approaches, see Mortensen 1998).

First, there is a difference in the assumption about the bargaining power of workers. In monopsony models, it is assumed that employers set wages unilaterally whereas the matching models typically assume some process of wage bargaining between employer and worker (although one could set up these models so that employers have all the bargaining power).[9]

[9] Adam Smith (1976, p. 84) had something to say about the practice of economists to see bargaining power of workers everywhere: "we rarely hear, it has been said, of the combinations of masters; though frequently of those of workmen. But whoever imagines, upon this account, that masters rarely combine, is as ignorant of the world as of the subject."

Secondly, there is a difference in the assumption made about the timing of wage determination. In the formal models of monopsony introduced in the next chapter, wages are modeled as being determined prior to an employer and a worker meeting each other: this is often called ex ante wage posting. In contrast, matching models typically assume that wages are determined after employer and worker have met (this is often called ex post wage bargaining).

If one judges theories by the realism of their assumptions, then I believe that the wage-posting monopsony model is to be preferred. This is not because it is the best description of the labor market in all circumstances (wage bargaining between employers and workers is observed), just that it is a better description most of the time. For example, chapter 5 documents the existence of a substantial number of firms (in labor markets without minimum wages or trade unions) that pay all their workers in a particular job the same wage. It is hard to see how this could be the outcome of individualized ex post wage bargaining between employers and workers given the heterogeneity of workers within the firms. Even in labor markets that one thinks of as being highly individualistic such as Wall Street, employers seem reluctant to engage in more than limited negotiation: Lewis (1989: 149) describes how Salomon Brothers lost their most profitable bond trader because of their refusal to break a company policy capping the salary they would pay. Models of wage posting seem to provide a better description of reality.

But, economists often also judge theories not by the realism of their assumptions but by the quality of their predictions. Comparing wage posting and wage bargaining models on this basis is difficult because so many of the predictions are the same and it may not matter greatly which assumption about wage determination is used in many circumstances.[10] There is a good reason for this: even though monopsony models appear to give all the bargaining power to the employer, both monopsony and matching models predict that the rents of the employment relationship get shared between workers and employers. In monopsony models, workers get some share of the surplus as long as employers are not perfectly discriminating monopsonists (and chap-

[10] However, there are some substantive differences. Ex post wage bargaining implies that all efficient matches will be consummated whereas ex ante wage posting may result in some efficient matches failing to be consummated (e.g., an unemployed worker with a particularly high reservation wage may not want the job at the offered wage even though there is a higher wage at which both employer and worker would gain from a match). However, ex post wage negotiation may not be effective in motivating ex ante investments by employers or workers as there is no guarantee that the rents from these investments will not be appropriated. On the other hand, the commitment implied by ex ante wage posting may be better able to motivate investments.

ter 5 argues that there are good reasons why they cannot be). Assuming that firms set wages and are monopsonists, at least in a formal sense, should not be taken to imply that their share of any rents is necessarily large.[11]

Another advantage of the monopsony over the matching approach is that it is much easier to forge links with other parts of labor economics. Although the underlying model of the labor market with frictions may be relatively complicated with a lot of dynamics and value functions, one can often represent and understand the decision problem of the individual employer in the monopsony model in terms of the textbook static model of monopsony. In contrast, the matching models do not have a simple static textbook counterpart model and the use of these models has led to unfortunate parallel literatures in which the same labor market phenomenon is "explained" by both a matching model and a conventional static model without the fundamental similarity between them being recognized. From those who specialize in the analysis of matching models, one often hears the claim that "dynamic models are different" to justify this state of affairs: while there is some truth in this statement, it is much less true than they commonly think. And empirical labor economists often feel that there is little benefit in terms of understanding and a considerable cost in terms of analytical complexity from using a dynamic model and fall back on the familiar textbook model of perfect competition.

Hence, although one should think in terms of monopsony and matching models as being fundamentally similar models of the labor market, the monopsony model is a better description of the way labor markets work and makes it much easier to forge links with the rest of labor economics.

1.4 Antecedents

As has already been pointed out, a number of distinguished economists have seen labor markets as operating in the way described in this book and bits and pieces of modern labor economics are, implicitly or explicitly, analyses of monopsonistic or oligopsonistic labor markets. But there are two particular traditions that need to be singled out as being important influences on this work.

The first is the labor economics of the so-called neorealist or neoclassical revisionist labor economists (Kaufman 1988) who thrived in the

[11] Some might object to the use of the word monopsony in a situation in which workers get some or even most of the rents. But, consumers are strictly better off with electricity than without although most people would be content with the description of the utility as a monopolist. The use of the word "monopsony" is simply meant to refer to the fact that employers set wages.

United States in the late 1940s and the 1950s before being supplanted by economists who drew their inspiration from Hicks' *Theory of Wages* and from the Chicago school of thought. These economists like John Dunlop, Clark Kerr, Richard Lester, and Lloyd Reynolds had been brought up on neoclassical economics but felt that the competitive model gave a seriously inaccurate picture of how labor markets operated.

There were two main reasons why they arrived at this conclusion. First, studies of labor mobility seemed to show that workers were extremely reluctant to change jobs and hence that the mechanism which was imagined to enforce the competitive law of one wage was, in reality, much weaker than most labor economists imagined. One consequence of this was that the "market" did not dictate the wage an employer had to pay or face ruin: employers had, in fact, considerable discretion in the wage that they chose to pay. Further evidence of this was the considerable dispersion in wages found in labor markets defined very tightly in terms of occupation and area (Lester 1946, 1948; Reynolds 1946a,b, 1951; Slichter 1950; Dunlop 1957, amongst others). They were well aware of the possibility that such wage dispersion might be driven by differences in the non-wage aspects of jobs or differences in worker quality, or be only short-term (see, e.g., the discussion in Lester 1952: 487–88) but they arrived at the conclusion (often more by the exercise of judgment than firm evidence) that the wage dispersion was real and permanent. The practical experience of several of these economists in the work of the War Labor Board which set out to find *the* market wage for particular classes of labor and found only wage dispersion was particularly important in convincing them that the competitive model suffered from serious deficiencies.

These economists were actively discussing the supply curve of labor to the firm, the issue that is at the heart of this book. Reynolds (1946a: 390) wrote in a paper entitled *The Supply of Labor to the Firm* that "the view that labor-market imperfections result in a forward-rising supply curve of labor to the firm appears to have been first elaborated by Mrs. Robinson. This conclusion has made its way rapidly into the textbooks and seems well on the way to being generally accepted as a substitute for the horizontal supply curve of earlier days." It is hard to imagine a paper with this title in the journals of today let alone a statement along these lines. Bronfenbrenner (1956: 578) wrote

> the typical employer in an unorganized labor market is by no means a pure competitor facing market wages which he cannot alter. The mobility of the labor force, even between firms located close together, is low by reason of the inability of workers to wait for employment or risk unemployment, plus the inadequacy of the information usually available to them regarding alternative

> employment opportunities. This low mobility permits each employer to set his own rates and form his own labour market within limits which at some times may be quite wide. In the technical jargon of economic theory, the typical employer in an unorganized labor market has some degree of monopsony power and can set his own wage policy

a statement of the central themes of this book which would be hard to better.

So these economists were writing about the issues on which I write and thinking about explanations along the same sort of lines. Yet, I cannot help feeling that these labor economists would not necessarily welcome my embrace.[12] My bald assumption that employers set wages to maximize profits is the kind of crass generalization from which someone like Lester instinctively recoiled. He came to emphasize how the lack of cutthroat competition in the labor market gave leeway for employers to pursue many ends and this was, for example, one explanation of wage dispersion observed (see, e.g., Lester 1952). Perhaps this was because he saw any model based on a single objective (like profit maximization) predicting a determinate outcome, a prediction that was then obviously falsified by observation of the world. But, the general equilibrium models that are used (see, e.g., the model of section 2.4) have as an equilibrium a range of wages even when the objectives pursued by all firms are identical: in a sense they are models of determinate indeterminacy.[13]

While Reynolds wrote about the supply curve of labor to the firm, his final conclusion was that "in actuality, an employer can usually expand and contract employment at will without altering his terms of employment" (Reynolds 1951: 227) so that the competitive labor supply curve gave the right answer though for the wrong reasons. He arrived at this conclusion primarily because of the observation that it did not seem to cost much in terms of time or money to recruit extra workers: that is, vacancy durations were (and are) extremely low. This is a serious objection to the relevance of the monopsony model and one which is discussed at length in chapter 10. But my conclusion is different: I argue that what we know about vacancies is perfectly consistent with the existence of non-negligible monopsony power.

The other important inspiration for this book is a single paper: Burdett and Mortensen (1998).[14] This paper was presented at the

[12] If one looks at the representative quotes about monopsony in the textbooks authored by these economists, it would be hard to see any more favorable inclination to monopsony than is found in the others.

[13] Though Lester's position does receive some support if our basic model is tweaked to introduce mobility costs and preferences over non-wage job attributes when multiple equilibria tend to be rife.

[14] It may have been published in 1998 but was originally written at least 10 years earlier.

LSE in 1990 and it was a revelation to me. Here was a simple elegant analytical framework that could explain the existence of equilibrium wage dispersion (and other stylized facts about the labor market). If I had not been quite so ignorant I would have realized that proving the possibility of equilibrium price or wage dispersion was not as new or as difficult as I had imagined (one might cite Butters 1977; Salop and Stiglitz 1977; Reinganum 1979; Burdett and Judd 1983; Albrecht and Axell 1984; Lang 1991; Montgomery 1991a, among others which did more or less the same thing). But the advantage of the Burdett and Mortensen model to me was that, whereas many of the other models of price or wage dispersion were too stylized to be able to take to labor market data, their model was expressed in terms of quit and recruitment rates, and job offer arrival rates that had obvious empirical counterparts. So it is their model that forms the basis of much of what follows, though I imagine that one could have built much of it on some of the other papers.

1.5 Summary of Chapters and Main Results

This book is based on two assumptions:

- there are important frictions in the labor market;
- employers set wages.

The consequence of the first assumption is that the employers have market power in the labor market and the consequence of the second is that they exercise it. The labor supply curve facing employers is not infinitely elastic so that they have some monopsony power. The style of this book is to systematically apply these two assumptions to most areas of labor economics.

The book is divided into four parts. In the first part, chapters 2 through 4, some basic models and results are laid out. Each chapter presents both the relevant theory and empirical evidence based on US and UK data. Every attempt has been made to make the main body of each chapter as accessible as possible with the proofs of the propositions and more technical material confined to an appendix at the end of each chapter. And, because the same data sets are used throughout, there is also a Data Sets Appendix at the end of the book, providing details of how the data were constructed.

Chapter 2, *Simple Models of Monopsony and Oligopsony* starts by presenting some partial equilibrium models of static and dynamic monopsony. While these partial equilibrium models are adequate for analyzing many questions, there are others for which it is necessary to model inter-

actions between employers, that is, to model the labor market as oligopsonistic. The chapter then presents a model of oligopsony based on the wage-posting model proposed by Burdett and Mortensen (1998).

The chapter derives the well-known result that the extent of employer monopsony power is related to the wage elasticity in the labor supply curve facing an individual employer: the less elastic the supply curve, the more market power the employer possesses. It also argues that the greater the ability of workers to move from employer to employer, the more wages will be driven up towards their marginal product. It suggests that the proportion of workers recruited directly from other jobs is a good simple measure of the competitiveness of labor markets. For both US and UK data sets, this proportion is shown to be in the region of 45–50%, a level that is argued to suggest employers have substantial market power.

Chapter 3, *Efficiency in Oligopsonistic Labor Markets*, considers the welfare implications of oligopsonistic labor markets in variations on the model of Burdett and Mortensen (1998). Most of the book is about the positive implications of assuming that employers have market power over their workers. But while "monopsony" as used in this book should be interpreted as a technical term to describe the situation where the labor supply curve to the firm is not infinitely elastic, the term often has more emotive connotations and is sometimes taken to imply that, in some sense, wages are "too low." This is certainly true for the textbook analysis of a single monopsonist where, if the employer has market power, one can always find a binding minimum wage that raises employment and welfare. However, as chapter 3 shows, this simple conclusion breaks down once one moves beyond the case of the single monopsonist. The main conclusion of the chapter is that the free market equilibrium is generally not efficient but that interventions like the minimum wage may improve or worsen efficiency, depending on the particular model being considered. Hence the chapter concludes that theory alone can be of little use in evaluating policy.

The final section of chapter 3 presents a simple model of a "ghetto," emphasizing how, in labor markets with frictions, it is relatively simple to generate multiple equilibria and agglomeration effects. For example, residents of a neighborhood may not invest in human capital if they think there are no jobs in which to use them, while employers may not locate in an area in which the residents have low levels of human capital. In a market with frictions, there is no mechanism to ensure that an act of investment in human capital by an individual will bring forward the investment of physical capital to employ it.

Chapter 4, *The Elasticity of the Labor Supply Curve to an Individual Firm*, presents evidence on the wage elasticity of the labor supply curve to the individual employer. This is the natural place to start to make the case

that monopsony is empirically relevant as the assumption that the labor supply curve to individual employers is not perfectly elastic is the fundamental idea in monopsony. There are astonishingly few papers in the labor economics literature on the supply of labor to individual employers in contrast to the volumes written about labor demand and individual labor supply to the market as a whole.

In estimating the supply curve to an individual employer, the obvious place to start is to regress log wages on the log of employment (plus other relevant controls). One finds, consistent with monopsony, a very robust positive correlation between wages and employment. This employer size–wage effect is well known in labor economics though it is rarely interpreted as evidence of an upward-sloping labor supply curve to an individual employer. The chapter reviews the more common explanations for the employer size-wage effect, concluding that none of them can explain it all, and that part of the employer size-wage effect does seem to be the result of an upward-sloping supply curve of labor to the individual employer. However, once one has controlled for other relevant factors, the elasticity of wages with respect to employment is often low, in the region of 0.04, implying that the elasticity in the labor supply curve to the employer is high—about 25. But, these OLS estimates are likely to be biased downwards because shifts in the supply of labor to the employer will tend to induce a negative correlation between wages and employment. Reverse regressions in which employment is regressed on wages suggest a much lower wage elasticity of the labor supply curve—often in the range of 1.5–3.5. Finding a suitable instrumental variable is the obvious way to try to sharpen up these estimates but that is not an easy task as the instrument needs to be firm specific. The few studies that do take this approach suggest that labor supply to individual firms is relatively inelastic.

The second half of chapter 4 takes a different approach to estimating the labor supply elasticity, based more explicitly on a dynamic model of monopsony. As, in steady state, employment, N, is equal to the recruitment rate, R, divided by the separation rate, s ($N = R/s$), the wage elasticity of employment can be written as the wage elasticity of recruitment minus the wage elasticity of separations. There is a relatively large existing literature that estimates the sensitivity of separations to the wage but estimating the elasticity of recruits is more difficult. However, it is shown how in models of dynamic oligopsony there will be a tight relationship between the separation and recruitment elasticities. In the simplest model, they must be equal to each other and, in more complicated models, a weighted average must be equal. Using this approach the wage elasticity of the labor supply curve to an individual employer is estimated to be in the region of 0.75–1.5, that is, relatively low.

The second part of the book, chapters 5 through 8, is about how monopsony can help us towards a better understanding of the observed distribution of wages.

Chapter 5, *The Wage Policies of Employers*, discusses the incentives for an employer to pay different wages to identical workers, that is, to become a discriminating monopsonist, and the difficulties with doing so. For example, employers would like to be able to pay low wages to workers with low reservation wages but it may be very difficult to observe reservation wages. Employers are more likely to base wage discrimination on non-manipulable characteristics of the workers like job tenure and age. The chapter shows how there are incentives for employers to use seniority wage schedules in line with what is observed. However, it is argued that there are good theoretical reasons and empirical evidence to suggest that the ability to wage discriminate may be severely limited in practice.

Chapter 6, *Earnings and the Life Cycle*, examines the way in which earnings evolve over a working life. The human capital approach to this question emphasizes the way in which both general and specific human capital accumulate over a lifetime and empirical correlations of earnings with experience (or age) and job tenure are normally interpreted in the light of the human capital model. Section 6.1 starts by presenting evidence that there is something wrong with this way of interpreting earnings functions. For example, the earnings losses of displaced workers are increasing in the level of experience, something that should not happen if the returns to experience represent the returns to general human capital. Section 6.2 then shows that a substantial part of the observed cross-sectional returns to job tenure is the result of the bias caused by the fact that those in high-wage jobs are less likely to leave them.

Section 6.3 then introduces a job-shopping model as a way to explain correlations between wages, age, and job tenure even if the wage offer distribution does not depend on age and job tenure. For example, there may be a correlation of wages with age because older workers are more likely to have found the better-paying jobs (Burdett 1978). One can then explain why more experienced workers suffer larger wage losses after displacement as job loss causes a reduction in "search capital." And there may be a correlation of wages with job tenure as those who have been lucky enough to find a high-paying job are less likely to leave it. However, as section 6.3 makes clear, the correlations predicted by the search model are more complicated than this simple discussion suggests.

Section 6.4 then proposes a new framework for decomposing the life cycle profile of earnings into three components: the growth in earnings on the job, the costs of job loss, and the return to job mobility. It is shown

how the returns to job tenure as conventionally measured are a weighted average of the change in the costs of job loss and the returns to job mobility but that this mixes up two very different processes as job mobility is mostly voluntary on the part of workers, leading to wage gains, while job loss is involuntary, leading to wages losses.

The final two sections then present two applications of this approach: estimating the returns to job mobility and the decline in average earnings among older men. It is shown how the decline in earnings among older men is primarily the result of substantial rates and costs of job loss.

Chapter 7, *Gender Discrimination in Labor Markets*, discusses how monopsony can help us understand the gender pay gap. It is argued that the weaker attachment of women to the labor market can go some way towards explaining the gender pay gap even if there is no gender productivity gap. The reason is that women will find it harder to work their way into the better-paying jobs. Furthermore, evidence is presented that women are less motivated than men by money in choosing jobs so that the female labor market is likely to be more monopsonistic than the male. Section 7.4 presents evidence for this from responses to questions on the motivation for changing jobs and section 7.5 presents evidence that the returns to job mobility are lower for women than for men. Human capital explanations of the gender pay gap also emphasize the weaker attachment of women to the labor market as a source of the gender pay gap but argue that this results in lower productivity. Two pieces of evidence inconsistent with this view are presented: in section 7.7, it is shown how the returns to job tenure are, if anything, larger for women than for men while section 7.8 analyzes the impact of the 1970 UK Equal Pay Act that resulted in a large increase in female relative wages but had no impact on relative employment contrary to the predictions of the human capital model.

Chapter 8, *Employers and Wages*, considers the well-known empirical "puzzle" that employer characteristics are correlated with wages. In a competitive market these correlations should not exist (abstracting from compensating wage differentials that do not seem to be empirically that important) as the wage should be determined solely by the characteristics of workers. However, as shown in section 8.1, we would expect wages to vary with employer characteristics like size, productivity, and profitability if employers have some market power. The "puzzle" is simply what we would expect.

Sections 8.2 and 8.3 discuss the implications of monopsony for the estimation of compensating wage differentials. It is argued that the conventional approach to estimating the value of non-pecuniary aspects of jobs that is based on estimating earnings functions is flawed if employers have market power as there is no reason to believe that utility is

equalized across jobs in the labor market. In particular, there is good reason to think that utility will be lower in jobs with worse work conditions. An alternative approach to estimating the value of non-pecuniary benefits based on estimating separation functions is proposed and an application to estimating the disamenity associated with night work is presented. Section 8.4 discusses the likely effect of mandated benefits, intervention to regulate the non-wage conditions of work, for example, health and safety legislation, maximum hours legislation, etc. In a competitive labor market, it is often argued that such legislation is likely to be bad as it imposes an inefficient wage-benefit combination and may actually harm rather than help workers. However, it is shown that this is not necessarily the case if employers have some market power: regulation of non-wage aspects of jobs will make workers better off as long as the non-wage attribute is a "normal" good and the regulation is not too onerous.

Finally, section 8.5 applies the framework established earlier in the chapter to the analysis of hours of work. The determination of hours of work as considered in the labor supply literature is normally treated as a completely different subject from the analysis of other non-wage job attributes. But there is no good reason for this: given the level of earnings, higher hours increase output and reduce worker utility just like any other non-wage attribute. It is argued that, if employers have monopsony power, then workers are likely to be overworked in the sense of being forced to work more hours than they would like given their wage.

The third part of the book, chapters 9 through 11, is concerned with the "quantity" side of the labor market, the supply of and demand for labor, and the determinants of investment in human capital.

Chapter 9, *Unemployment Activity and Labor Supply*, considers the determinants of the level and structure of unemployment and inactivity from the perspective of the worker. The employment rate of individual workers is determined by the rate at which they get jobs when not in employment and lose them when in employment. The main way in which individuals can influence the rate at which they get jobs is by their choice of job search activity. Section 9.1 endogenizes the choice of search intensity both on and off the job. The relative effectiveness of these two types of job search is important and a new test is proposed based on the fact that the reservation wage should depend positively (negatively) on the productivity of workers as off-the-job search is more (less) effective than on-the-job search. This empirical evidence strongly suggests that off-the-job search is more effective. Section 9.2 then discusses the distinction between unemployment and inactivity as defined in labor market statistics. Competitive models of the labor market do not have a meaningful distinction between these two labor market states but because the unemployed are defined as those with job

search intensity above a critical level, the framework of this chapter makes the distinction easy to understand. An application to the discouraged worker effect suggests that, when aggregate labor market conditions worsen, job search intensity falls resulting in a rise in measured inactivity rates.

Section 9.3 considers the job search intensity of the employed. Monopsony has a strong prediction, that job search activity should be declining in the wage as there are then fewer opportunities to find a better job. The empirical evidence reported is strongly in support of this prediction. Section 9.4 then considers the determinants of the rate at which workers will quit jobs for non-employment. Consistent with the empirical evidence, the model predicts that quit rates will be declining in the wage.

Sections 9.6 and 9.7 are concerned with conceptual issues about the nature of unemployment in labor markets where employers have market power. In the simplest models of monopsony, unemployment appears "voluntary" in the sense that all employers would like to hire more workers at the going wage. This seems hard to reconcile with the observation that jobs often seem to be hard to find and the feeling that many economists have that unemployment is "involuntary." However, as sections 9.6 and 9.7 show, it is a simple matter to reconcile models of monopsony with models of involuntary unemployment (represented by efficiency wage models).

Chapter 10, *Vacancies and Labor Demand*, considers the determinants of the level of employment from the perspective of employers. Sections 10.1 and 10.2 are concerned with the interpretation of vacancy statistics. It is argued that, to have a meaningful model of vacancies, one has to have a model in which the creation of jobs requires some ex ante investment and in which the supply of labor to the firm is stochastic. With these features, a model of the labor market in which employers have considerable market power is quite consistent with the observation that vacancy rates are low, and vacancies are typically of short duration and have relatively small numbers of applicants. Empirical evidence supports the conclusion that those firms that pay higher wages have fewer difficulties in filling vacancies.

Sections 10.3 and 10.4 are concerned with the technology by which workers and employers are matched. A crucial issue turns out to be whether large employers have an intrinsic advantage over small firms in recruiting workers, because this is important in determining the wage elasticity in the supply of labor to the firm. However, it is shown that large employers are not more likely than small firms to use recruitment methods in which they might be thought to have an advantage, like social contacts.

Finally, section 10.6 contains a brief discussion of the determination of lay offs, arguing that there are good reasons to think that they will occur while there is still some surplus in the relationship remaining for workers.

Chapter 11, *Human Capital and Training*, considers the incentives for the acquisition of human capital in monopsonistic labor markets. Section 11.1 considers the incentives for workers to engage in the acquisition of education before they enter the labor market. Because part of the returns to any such education is likely to accrue to future employers of the worker, there is a prima facie case for believing there will be underinvestment in human capital. However, there is some reason to believe that the labor market for more educated workers may be less monopsonistic in which case it may be that this conclusion is misleading. Section 11.2 then considers the provision of employer-provided general training. A key prediction of the monopsony model which contrasts very strongly with that of the competitive model is that employers will be prepared to pay for some investments in general training because they can expect to get some returns from it. However, because future potential employers of a worker might also expect to get a share of the returns from any investment in human capital, one would expect to see underinvestment. Section 11.3 then considers firm-specific training. A striking conclusion is that workers may capture a higher share of the returns to firm-specific investments than of general investments if employers have market power. Section 11.4 concludes with a discussion of the empirical evidence on training.

The final part of the book, chapters 12 and 13, considers the impact of institutions that interfere with the ability of employers to set wages and draws some conclusions.

Chapter 12, *Minimum Wages and Trade Unions*, is concerned with the impact of these wage-setting institutions on wages and employment. Although these institutions are often seen as essentially similar (they both raise wages above the market-clearing level), their effects in a monopsonistic labor market are likely to be rather different. For example, minimum wages have a direct impact on the lowest wages in a given labor market so are likely to "push" the wage distribution from below, while trade unions are likely to set the highest wages in a given market so will "pull" the wage distribution from above. Section 12.1 discusses the impact of the minimum wage on the wage distribution. Empirical evidence is presented that spillover effects from the minimum wage onto the US wage distribution are substantial. Section 12.2 then argues that much of the evolution of wage inequality in the bottom half of the US wage distribution from 1980 to 2000 can be explained by variation in the minimum wage. Section 12.3 then discusses the controversial issue of the impact of the minimum wage on employment. While a minimum wage does not necessarily cost jobs in an oligopsonistic labor market, it is shown that

the simple result from the model of a single monopsonist, that a suitably chosen minimum wage must raise employment, does not carry over to a labor market in which one models interactions between firms and heterogeneity among them. An open-minded empirical approach is appropriate for investigating the impact of minimum wages on employment.

Section 12.4 discusses how models of trade unions need to be modified to recognize the fact that employers have some market power. It also discusses the argument that "corporatist" systems of wage bargaining can do something to alleviate the problems caused by a "free market" system of wage determination. Section 12.5 discusses the impact of trade unions on wages. It focuses on the impact of unions on non-union wages, arguing that in a labor market where employers have some power over wages, the impact of unions on non-union wages is likely to depend on whether an on- or off-the-job search is more effective. The evidence presented in chapter 9 suggests that an off-the-job search is more effective in which case unions would be expected to raise non-union wages. Empirical evidence for this is presented and it is argued that the correlations cannot be explained by the "threat" effect.

Chapter 13, *Monopsony and the Big Picture*, offers some conclusions. Section 13.1 reviews the sources of monopsony power and the evidence that employers have it. Section 13.2 argues that recognizing the existence of monopsony power in the labor market does not mean supplanting all existing competitive analysis: in many cases, it simply adds to it. One might wonder about how important monopsony is in understanding the "big" issues of the day. Section 13.3 addresses this argument by arguing that a view that the labor market is monopsonistic is necessary for an adequate understanding of changes in the bottom half of the US labor market since 1980. Section 13.4 then discusses what monopsony has to say about the design of labor market policy. The main substantive conclusion is that labor economists should be more open-minded about the likely impact of labor market interventions: empirical evidence is more powerful than theory. Too often (e.g., in discussions of European unemployment), labor economists simply assume (often unthinkingly) that the alternative to a regulated labor market is a labor market that is well approximated by the perfectly competitive model.

In the book as a whole, virtually all of the main topics of labor economics are covered although not necessarily in a familiar order. Table 1.2 presents a simple key to where some topics may be found in this book.

TABLE 1.2
Topics in Labor Market Analysis

Traditional Subject	*Location in this Book*
Labor supply (hours)	Chapter 8
Labor supply (participation)	Chapter 9
Labor demand	Chapter 10
Compensating wage differentials	Chapter 8
Employers and wages	Chapter 8
Gender discrimination	Chapter 7
Earnings functions	Chapter 6
Employment contracts	Chapter 5
Efficiency wages	Chapter 9
Rent sharing	Chapter 8
Employer-size wage effect	Chapter 4
Human capital	Chapter 11
Minimum wages	Chapter 12
Trade unions	Chapter 12

2

Simple Models of Monopsony and Oligopsony

This chapter introduces some simple models of monopsony and oligopsony which form the foundation for the analysis in the rest of the book. The first three sections present some partial equilibrium models: the textbook static model of monopsony, a simple model of dynamic monopsony, and what is called a generalized model of monopsony where the firm has instruments other than the wage to influence the flow of recruits. The fourth section then presents a general equilibrium model of dynamic oligopsony (based on a simplified version of Burdett and Mortensen, 1998) to show how the framework is a fully coherent vision of the labor market as a whole. Although this model is highly stylized, it does capture the most important features of a labor market with frictions. Workers are faced with a distribution of wages so that there are good jobs and bad jobs. They try to get themselves into the good jobs but their progress resembles a game of "snakes and ladders." Sometimes they meet a "snake" and suffer the misfortune of losing their job and sometimes they find a "ladder" and have the good fortune to move to a better job. From the perspective of employers, the frictions in labor markets give them some discretion in setting wages. If they lower wages, they find it more difficult to recruit and retain workers but the existing workers do not all leave immediately and they continue to be able to recruit some workers so that they retain some workers even in the long run. The wages that employers set are influenced by competition from other employers but this competition is neither so cutthroat as to enable workers to extract all the surplus from the employment relationship, nor so weak as to enable employers to extract all the rents.

The chapter concludes by arguing that the fraction of recruits from non-employment is a good "back-of-the-envelope" measure of the extent to which workers are able to freely move between employers and, hence, of competition among employers for workers and the extent of market power possessed by employers in the labor market. Empirical evidence from the United Kingdom and the United States suggests that 45–55% of recruits were previously non-employed, a level which is likely to give employers considerable market power.

2.1 Static Partial Equilibrium Models of Monopsony

Given the lack of attention paid to monopsony in much of labor economics it is perhaps helpful to start with a quick review of the static textbook model of monopsony. In this model, the firm is assumed to face a labor supply curve that relates the wage paid, w, to the level of employment, N. Denote the supply of labor to the firm if it pays w by $N(w)$. Also, denote the inverse of this relationship by $w(N)$. Both $N(w)$ and $w(N)$ will be referred to as the labor supply curve to the individual firm. Total labor costs are given by $w(N)N$. Assume that the firm is a simple monopsonist who has to pay a single wage to all its workers (the incentives for wage discrimination are discussed in chapter 5). Assume the firm has a revenue function $Y(N)$. It wants to choose N to maximize profits which are given by

$$\pi = Y(N) - w(N)N \tag{2.1}$$

This leads to the first-order condition

$$Y'(N) = w(N) + w'(N)N \tag{2.2}$$

The left-hand side of (2.2) is the marginal revenue product of labor. The right-hand side is the marginal cost of labor, the increase in total labor costs when an extra worker is hired. The marginal cost of labor has two parts: the wage, w, that must be paid to the new worker hired and the increase in wages that must be paid to all existing workers. The solution is represented graphically in figure 2.1. Equilibrium is on the labor supply curve with the wage paid to workers being less than their marginal revenue product. Although the employer is making positive profit on the marginal worker, there is no incentive to increase employment because doing so would require increasing the wage (to attract the extra worker) and this higher wage must be paid not just to the new worker but also to all the existing workers. One particularly useful way of representing the choice of the firm is that marginal cost of labor is a mark-up on the wage, the mark-up being given by the elasticity of the labor supply curve facing the firm. Let us write the elasticity of the labor supply curve facing the firm as $\varepsilon_{Nw} = wN'(w)/N(w)$and let ε be the inverse of this elasticity. Then (2.2) can be written as

$$\frac{Y' - w}{w} = \frac{1}{\varepsilon_{Nw}} = \varepsilon \tag{2.3}$$

so that the proportional gap between the wage and the marginal revenue product is a function of the elasticity of the labor supply curve facing the firm. The gap between the wage and the marginal revenue product is what Pigou (1924) and Hicks (1932) referred to as the rate of exploitation and

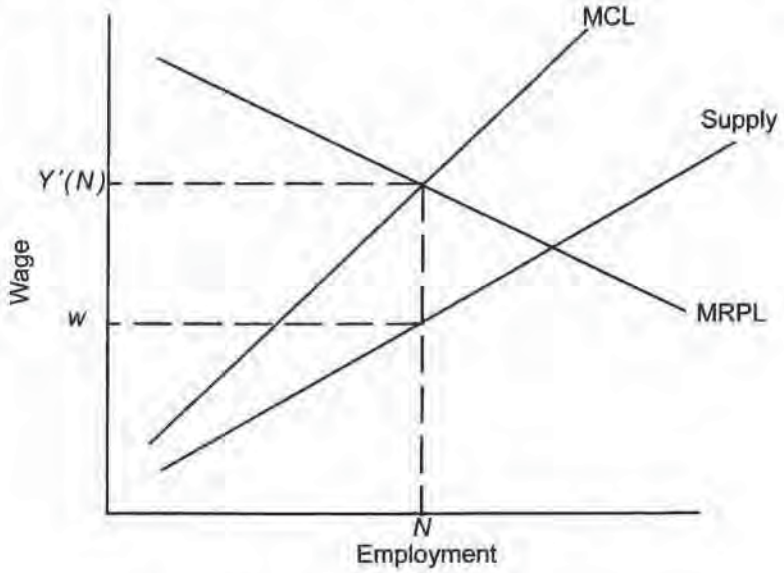


Figure 2.1 The textbook model of monopsony.

we will follow this tradition. Perfect competition corresponds to the case where $\varepsilon_{Nw} = \infty$ and $\varepsilon = 0$ in which case (2.3) says that the wage will be equal to the marginal revenue product.

Some of the comparative statics of the static monopsony model are the same as in the competitive model and some are different. For example, an increase in the marginal revenue product of labor will lead to an increase in employment and a rise in wages. The former would occur in a competitive model but the latter would not as a competitive firm would simply continue to pay the market wage. The impact of shifts in the labor supply curve to the firm are more complicated as the impact depends on how the change affects the marginal cost of labor and not just the average cost of labor. An increase in the supply of labor to the firm that keeps the elasticity the same will result in a rise in employment and a fall in wages just as in the competitive model. But, matters are more complicated if the elasticity of the labor supply curve changes as the average and marginal cost of labor can move in opposite directions; the most familiar example of this is the impact of a minimum wage. The minimum wage raises the average cost of labor but (if it is binding) reduces $w'(N)$ so its effect on the marginal cost of labor (see (2.2)) is ambiguous. In fact, one can show that a minimum wage that just binds must raise employment (a demonstration of this can be found in most labor economics textbooks).

2.2 A Simple Model of Dynamic Monopsony

One might wonder how this completely static model of the labor market corresponds to the description of the labor market in the first chapter that was based on dynamic arguments. The static and dynamic models can be linked in the following way. Assume that workers leave the firm at a rate s that depends negatively on the wage paid, and recruits arrive at the firm at a rate R that depends positively on the wage. If the firm had N_{t-1} workers last period and pays w_t this period, its labor supply will be

$$N_t = [1 - s(w_t)]N_{t-1} + R(w_t) \tag{2.4}$$

where $s(w)$ is the separation rate and $R(w)$ the recruitment rate.

In a steady state, total separations sN must equal recruits R so that we have

$$N(w) = \frac{R(w)}{s(w)} \tag{2.5}$$

giving us a positive long-run relationship between employment and the wage.[1] In this case the elasticity of the labor supply curve facing the firm can be written as

$$\varepsilon_{Nw} = \varepsilon_{Rw} - \varepsilon_{sw} \tag{2.6}$$

where ε_{Rw} is the elasticity of recruits with respect to the wage and ε_{sw} is the elasticity of separations with respect to the wage.

In a dynamic model, there is an important distinction between the elasticity of the short-run labor supply curve facing the employer and the long-run elasticity. The elasticity of (2.6) is the long-run elasticity of the labor supply curve facing the firm. The short-run elasticity, denoted by ε^s_{Nw}, is the elasticity of N_t with respect to w_t holding N_{t-1} fixed. Differentiating (2.4), we have

$$\begin{aligned}\varepsilon^s_{Nw} &= \frac{w_t}{N_t}\frac{\partial N_t}{\partial w_t} = -w_t s'(w_t)\frac{N_{t-1}}{N_t} + \frac{w_t R'(w_t)}{N_t} \\ &= -\varepsilon_{sw}s(w_t)\frac{N_{t-1}}{N_t} + \varepsilon_{Rw}\frac{R(w_t)}{N_t} \\ &= s(w_t)[\varepsilon_{Rw} - \varepsilon_{sw}]\frac{N_{t-1}}{N_t} + \varepsilon_{Rw}\frac{N_t - N_{t-1}}{N_t}\end{aligned} \tag{2.7}$$

In a steady state in which $N_t = N_{t-1}$, the elasticity of the short-run labor supply curve facing the firm, ε^s_{Nw}, can, using (2.6), be written as

[1] One can invert $N(w)$ to give $w(N)$, the wage the firm must pay if it wants to have N workers.

$$\varepsilon^{s}_{Nw} = s(w_t)\varepsilon_{Nw} \tag{2.8}$$

(2.8) shows that the short-run labor supply curve facing the firm is less elastic than the long-run one (as $s(w_t) < 1$), and the difference is greater the lower the separation rate.

In a dynamic monopsony model, it is not immediately clear whether the short- or long-run labor supply elasticity is most relevant for wage determination. If firms must commit to a particular wage and do not discount the future, they will be interested in maximizing steady-state profits and the formula in (2.3) will still hold where the relevant elasticity is the long-run one. But, suppose firms do discount the future at a rate D and cannot make long-term commitments on the wage. In particular, suppose the firm cannot commit itself to a particular wage for more than a single period in advance so that the promise of a particular wage this period carries no guarantee that it will be paid next period.[2] The following result (Boal and Ransom 1997) tells us about the steady-state relationship between the marginal revenue product of labor and the wage in this case.

Proposition 2.1. *In a steady state, the relationship between the marginal revenue product and the wage is given by*

$$\frac{Y'(N) - w}{w} = \varepsilon\left[1 + \frac{(1-D)(1-s)}{s}\right] = (1-D)\varepsilon^{s} + D\varepsilon \tag{2.9}$$

where ε is the inverse of long-run elasticity of the labor supply curve and ε^{s} is the inverse of the short-run elasticity.

Proof. See Appendix 2.

(2.9) says that the rate of exploitation is a weighted average of the long-run and short-run elasticities with the weight on the long-run elasticity being the discount factor. The more employers discount the future, the greater the weight given to the short-run elasticity and the larger will be the rate of exploitation (as it will be the case that $\varepsilon^{s} > \varepsilon$). The intuition is that cutting wages is more attractive when employers discount the future more heavily as the costs of this strategy (lower future labor supply) do not weigh so heavily on their minds.

One variant of (2.9) is where the length of the "period" (of wage commitment) goes to zero. If the length of period is Δ, and d and ς are the instantaneous interest and separation rates, respectively, we will have

[2] It is convenient to work in discrete time although we will consider the limit as the length of time between periods goes to zero.

$D = e^{-d\Delta}$ and $s = 1 - e^{-s\Delta}$. Then taking the limit as $\Delta \to 0$, we have

$$\frac{Y'(N) - w}{w} = \varepsilon\left[1 + \frac{d}{s}\right] \tag{2.10}$$

The difference between the right-hand sides in (2.3) and (2.10) is probably rather small for plausible values of d and s (perhaps an annual interest rate of 5% and 20% for the labor turnover rate) although (2.10) does suggest a larger rate of exploitation than (2.3).

In the interests of simplicity, most of the theoretical analysis in this book is based on the assumption that employers do not discount the future and choose wages once-for-all. In this case, it is the long-run labor supply elasticity that is important and, as this section has demonstrated, this is likely to understate the true extent of monopsony power.

2.3 A Generalized Model of Monopsony

In the models of monopsony considered so far, there is only one way for a firm to get employment of N and that is to pay the wage $w(N)$. In reality, firms can influence their employment through other means, for example, by varying the intensity of their recruitment activity. In this section we present a simple, yet general and flexible framework for thinking about monopsony in this situation.

Define the labor cost function, which we denote by $C(w, N)$, as the cost per worker, excluding direct wage costs, of keeping employment at N when the firm pays a wage w.[3] Some examples might make the idea clearer. For example, if recruiting and training a worker costs T (independent of the number of recruits) and the separation rate is $s(w)$, a flow of sN recruits is needed to maintain employment at N so that $C(w, N) = T/s(w)$. In this case, the labor cost function is independent of N. But, if it becomes increasingly hard to recruit and train workers

[3] One can think of both perfect competition and the static model of monopsony as being particular forms (albeit, non-differentiable) of the labor cost function. The traditional static monopsony model implicitly assumes that a firm that pays a wage w incurs no recruitment costs if it wants employment less than the labor supply forthcoming at that wage, but that there is no way at all for the firm to attract more workers. So, the form of the labor cost function in this model is $C(w, N) = 0$ if $N < N(w)$ and $C(w, N) = \infty$ if $N > N(w)$. The labor cost function for the competitive labor market model is the following. If there are no recruitment/training costs then, if w^c is the competitive wage, the labor cost function for the competitive model can be thought of as $C(w, N) = 0$ if $w \geq w^c$, and $C(w, N) = \infty$ if $w < w^c$. This says that any amount of labor can be recruited at zero cost as long as the wage paid is at or above the competitive level, but that no labor is available at any cost if a wage below the competitive wage is offered.

then the cost of recruiting and training workers $T(R)$ will be an increasing function of R and the labor cost function will take the form $C(w, N) = T(N/s(w))/s(w)$ in which case it will depend positively on employment. The issue of whether there are diseconomies of scale in recruitment and training turns out to be of some importance.

Now consider the optimal choice of the wage and employment. If we assume the firm has a revenue function $Y(N)$, steady-state profits can be written as

$$\pi = Y(N) - [w + C(w, N)]N \tag{2.11}$$

A more sophisticated analysis would recognize that recruitment takes time and there is a need to pay attention to the date at which costs are incurred and revenues accrue but the decision problem at the end of the day can normally be written as something that looks like (2.11) (for an explicit justification of this claim, see Manning 2001a).

A difference from the basic monopsony model is that the firm has a choice of the wage it can pay if it wants to maintain employment at N so that both wages and employment are choice variables that can vary independently of each other. Given N, it is optimal for the firm to choose w to minimize direct and indirect labor costs so let us define the function $\omega(N)$ as

$$\omega(N) = \min_{w} w + C(w, N) \tag{2.12}$$

Profits can then be written as

$$\pi = Y(N) - \omega(N)N \tag{2.13}$$

A comparison of (2.1) and (2.13) should make apparent the relationship between the model presented here and the basic monopsony model: it is that the labor supply curve $w(N)$ needs to be replaced by the labor supply curve $\omega(N)$. As $\omega(N)$ is the relevant labor supply curve, let us call it the effective labor supply curve. We can represent the decision problem for the employer as in figure 2.1 with $w(N)$ replaced by $\omega(N)$, the effective supply of labor to the firm. Unsurprisingly, it is going to be of some interest whether $\omega(N)$ is increasing in N which would give us the equivalent of an upward-sloping labor supply curve.

By application of the envelope theorem to (2.12), we have

$$\omega'(N) = C_N(w(N), N) \tag{2.14}$$

where $w(N)$ is the wage chosen if employment is N. Hence, the effective labor supply curve facing the firm is upward-sloping if the labor cost function is increasing in employment, that is, if there are diseconomies of scale in recruitment and training. If the level of employment has no impact on recruitment and training costs, then the effective labor supply

curve facing the firm will be infinitely elastic and will resemble the labor supply curve in a perfectly competitive market. Given this discussion, it should be apparent that the form of the labor cost function $C(w, N)$ is of some importance.[4] The labor market is "monopsonistic" if $C_N > 0$ so that the non-wage costs are increasing in employment and "competitive" if $C_N = 0$.

There is a reasonable argument that the labor cost function $C(w, N)$ should be used in all the analysis that follows, and that analysis suggests we should focus on the effective labor supply function $\omega(N)$ rather than the labor supply function $w(N)$. However, this is hard to do as we rarely have the requisite data on non-wage labor costs like training and recruitment costs. However, in chapter 10 we present some evidence that $C_N > 0$ so that there are diseconomies of scale in recruitment activity.

All of the models considered so far in this chapter have been partial equilibrium models in which the influence of factors external to the individual firm have been buried in its labor supply curve, $N(w)$. One could introduce general equilibrium considerations by explicitly allowing the actions of other firms to affect the supply of labor to the firm (or the labor cost function). But, such an approach would inevitably be ad hoc and it is best to construct an explicit general equilibrium model of an oligopsonistic labor market to check that the model as a whole is internally consistent. This is the subject of the next section.

2.4 A General Equilibrium Model of Oligopsony

Any general equilibrium model of oligopsony must model interactions between employers in an internally consistent manner. There are a number of ways in which this has been done in the literature: for example, Bhaskar and To (1999) use a Hotelling-style location model. Here we outline another such model developed by Burdett and Mortensen (1998) which can be thought of as a general equilibrium version of the dynamic monopsony model described in section 2.2. The assumptions made about the labor market are the following.[5]

(A1) Workers: There are M_w workers all of whom are equally productive and attach equal value, b, to leisure.

(A2) Employers: There are M_f employers, each of which is assumed to

[4] More detailed analysis of the comparative statics of the generalized model of monopsony can be found in Manning (2001a).

[5] These assumptions have been chosen to be the simplest possible whilst retaining the essential features of an oligopsonistic labor market. Many of these assumptions are relaxed at various points in the book or in other papers in the literature.

be infinitesimally small in relation to the market as a whole.[6] All employers have constant returns to scale, the productivity of each worker being p. For future use, denote the ratio of firms to workers by $M = M_f/M_w$.

(A3) Wage-setting: Employers set wages once-for-all to maximize steady-state profits (which is equivalent to assuming there is no discounting). All workers within a firm must be paid the same wage. Denote the cumulative density function of wages across employers by $F(w)$ and the associated density function by $f(w)$.

(A4) Matching Technology: Both employed and non-employed workers receive job offers at a rate λ. Job offers are drawn at random from the set of firms, that is, from the distribution $F(w)$. Employed workers leave their jobs for non-employment at an exogenous job destruction rate δ_u. All workers, both employed and non-employed, leave the labor market at a rate δ_r, to be replaced by an equal number of workers who initially enter non-employment. For future use, define $S = S_u + S_r$.

These assumptions are simpler than those used in Burdett and Mortensen (1998) but capture the essence of their model. Now, consider the equilibrium in the basic model.

The Behavior of Workers

The behavior of workers in this labor market is very simple. An employed worker will move to another job whenever a wage offer above the current wage is received. A non-employed worker will accept a job whenever the wage offer received is above some reservation wage, r. As job offers arrive at the same rate whether employed or non-employed, the decision to accept a current job offer has no consequences for future job opportunities. So, the job will be taken if it makes a worker better off now than they would be if non-employed, that is, if the wage exceeds the value of leisure. Hence, the reservation wage, the lowest wage for which workers will be prepared to work will simply be equal to b, the value of leisure. Later, in chapter 9, we analyze the determinants of the reservation wage in a more complicated setup where on- and off-the-job searches differ in their effectiveness.

The Employer's Decision

The employer's decision in this model is to choose the wage to maximize profits $\pi = (p - w)N(w; F)$ where $N(w; F)$ is the steady-state level of employment in a firm that pays a wage w when the distribution of wages in the market as a whole is F. So, prior to considering the profit

[6] Note that this assumption implies that employers do not have to be "large" to possess market power.

maximization decision, we need to consider employment determination.

Employment Determination

An employer who pays a wage w will recruit workers from among the non-employed (as long as w is larger than the reservation wage b) and from workers in other firms that pay less than w. The employer will lose workers who exit to non-employment or leave the labor force or who quit to other firms that pay higher wages. In general terms, if $s(w; F)$ is the separation rate and $R(w; F)$ is the recruitment rate, we must have in a steady state that

$$s(w; F)N(w; F) = R(w; F) \tag{2.15}$$

so that $N(w; F)$ is the level of employment at which the flow of recruits equals the flow of separations. In deriving $N(w; F)$, a very useful result is the following.

Proposition 2.2. *If $\infty > \lambda/\delta > 0$, the equilibrium must be a distribution of wages without any spikes.*

Proof. See Appendix 2.

The result that the equilibrium of this model must have wage dispersion even though all agents (both workers and firms) are assumed identical is the most striking feature of Burdett and Mortensen (1998). The intuition for it is not that easy to understand but the result comes from the fact that if there is a wage paid by a non-negligible fraction of employers, then paying an infinitesimally higher wage means that the employer starts to recruit workers from these employers leading to a discontinuous jump in the number of workers but only an infinitesimal fall in profits per worker. Hence, profits must rise and the initial situation could not have been in equilibrium.

The proposition implies that the equilibrium outcome must be a wage distribution with a continuous cumulative density function, $F(w)$. As all firms are identical but, in equilibrium, choose different wages which yield the same level of profit, there is an indeterminacy in equilibrium in the sense that which firms choose which wages is not defined and one might think there is a potential problem in ensuring that the right distribution of wages results from the uncoordinated choices of firms. This is a common problem in much of economic theory where the equilibrium involves mixed strategies. But, it is not a real problem here. The smallest differences in firms will result in a fully determinate equilibrium (see chapter 8).

As it is reasonable to believe that firm heterogeneity exists, the model presented here should be thought of as the limiting equilibrium as firm heterogeneity disappears.

From the analytical point of view this proposition is extremely convenient as it means that we can restrict attention to wage distributions $F(w)$ that are continuous. From a more practical point of view, the result has both advantages and disadvantages. The advantage is that the model can explain the existence of equilibrium wage dispersion, the well-documented fact that equally productive workers receive different wages according to who they work for (see, e.g., Lester 1946, 1952; Slichter 1950; Reynolds 1951; Dunlop 1957; Krueger and Summers 1988; amongst others). The disadvantage is that we do observe concentrations of workers (or "spikes") at particular points in wage distributions, often at the minimum wage or at "round" numbers.[7]

Now, consider how we can derive the supply of labor to a firm who pays a wage w. From (2.15) it is helpful to derive the separation and recruitment rate separately. The separation rate in a firm that pays w is

$$s(w; F) = \delta + \lambda[1 - F(w)] \tag{2.16}$$

as workers leave for non-employment at a rate δ, receive other job offers at a rate λ and a fraction $[1 - F(w)]$ of these offers are better than their current wage.

Deriving the flow of recruits to the firm, $R(w; F)$, is slightly more complicated. It is helpful to first derive the non-employment rate and the distribution of wages across workers.

The Non-Employment Rate

The non-employment rate, u, is simply given by

$$u = \frac{\delta}{\delta + \lambda} \tag{2.17}$$

as workers leave employment for non-employment at a rate δ and obtain jobs at a rate λ.

The Distribution of Wages Across Workers

The distribution of wages across firms is denoted by $F(w)$. This is not the same as the distribution of wages across workers as the systematic search by workers for better-paying jobs means that they will be concentrated in

[7] Modifying the model to allow for the existence of spikes (e.g., because of mobility costs) is likely to increase the monopsonistic elements in the model so moves us even further away from the competitive model than the current framework.

higher-wage firms. Denote by $G(w; F)$ the fraction of employed workers receiving a wage w or less when the wage offer distribution is F. The following proposition shows that there is a simple relationship between G and F.

Proposition 2.3. *The fraction of workers in employment receiving a wage w or less is given by*

$$G(w; F) = \frac{\delta F(w)}{\delta + \lambda[1 - F(w)]} \tag{2.18}$$

Proof. See Appendix 2.

From inspection of (2.18) one can see that $G(w; F) < F$ for $0 < F < 1$ so that workers are concentrated in the better-paying jobs, implying that such firms must have a higher level of employment. This is easy to understand: higher-wage firms have lower separation rates and higher recruitment rates so that they have more workers in a steady state. Some special cases may help the understanding of (2.18): as $\lambda \to 0$ so that opportunities to move up the job ladder once in employment are reduced, then $G(w; F) \to F(w)$ so the distribution of wages across workers converges to the distribution of wages across firms. On the other hand, as $\lambda \to \infty$ so that opportunities to move up the job ladder once in employment come at a very fast rate then $G(w; F) \to 0$ if $F(w) < 1$ and $G(w; F) \to 1$ if $F(w) = 1$ so that all workers end up in the firm that pays the highest wage.

The Flow of Recruits to a Firm

Now let us go back to deriving the level of employment in a firm that pays w. Recruits to this firm will come from non-employment and those employed in lower-wage jobs. There are $\lambda u M_w$ non-employed workers who receive job offers which are shared equally over the M_f firms so that the flow of non-employed recruits to the firm will be $\lambda u M_w / M_f = \lambda u / M$. Similarly, there are $\lambda(1 - u)G(w; F)M_w$ workers currently earning less than w who get job offers which again are spread over the M_f firms. So, the total flow of recruits to a firm that pays w is given by

$$R(w; F) = \frac{\lambda}{M}[u + (1 - u)G(w; F)] = \frac{\delta\lambda}{M[\delta + \lambda(1 - F(w))]} \tag{2.19}$$

where the second equality follows from use of (2.17) and (2.18). Combining (2.15), (2.16), and (2.19), we finally have the following expression:

$$N(w; F) = \frac{\delta\lambda}{M[\delta + \lambda(1 - F(w))]^2} \tag{2.20}$$

for the supply of labor to the firm. This captures the most important idea in the analysis of monopsonistic labor markets, namely that the labor supply to an individual firm is increasing in the wage paid so that the labor supply curve facing an individual firm is not infinitely elastic as is assumed in perfect competition. Employment is increasing in the wage because the separation rate is decreasing in the wage (a higher wage means workers are less likely to get a better job offer) and the flow of recruits is increasing in the wage (a higher wage means there are more workers in lower-wage firms).

The wages paid by other firms are also important in determining the supply of workers to a firm. In fact, in (2.20) the position of the firm in the wage offer distribution is a sufficient statistic for the supply of labor to the firm. This is not true in more general models but one should still think of the supply of labor to the firm as being determined by the wage offered relative to the alternatives of non-employment or employment in other firms.

The Employer's Decision Revisited

Given (2.20), profits can be written as

$$\pi(w; F) = \frac{\delta\lambda(p - w)}{M[\delta + \lambda(1 - F(w))]^2} \tag{2.21}$$

Every firm will choose its wage to maximize profits.

Equilibrium

We need to find an equilibrium wage distribution $F(w)$. $F(w)$ will be an equilibrium if two conditions are satisfied:

- all wages that are offered yield the same level of profits;
- no other wage yields a higher level of profits than a wage that is offered.

For the special model considered here, one can (as Burdett and Mortensen 1998 showed) derive a closed-form expression for the equilibrium wage distribution w. The easiest way to derive this equilibrium is in stages.

Proposition 2.4. *The lowest wage offered in equilibrium is the reservation wage, b.*

Proof. See Appendix 2.

The intuition for this result is simple. There is no point in an employer paying a wage lower than b as no workers will accept such a low wage offer and employment and profits will be zero. And there is no point in the lowest-wage employer paying a wage strictly above b as, from (2.20), the supply of labor to the firm is a function of its position in the wage distribution (F) and not the actual wage paid. So, cutting the wage to b will lead to the same level of employment but higher profits per worker.

Given that the lowest wage offered is b (which is also the reservation wage), the equilibrium level of profits, π^*, can be found by using this fact in (2.21) to give

$$\pi^* = \frac{\delta\lambda(p-b)}{M[\delta + \lambda]^2} \tag{2.22}$$

The equilibrium wage distribution $F(w)$ can then be found by equating (2.21) to (2.22). After some re-arrangement this leads to the following.

Proposition 2.5. *The offered wages lie in the interval*

$$b \leq w \leq p - \left(\frac{\delta}{\delta+\lambda}\right)^2 (p-b) \tag{2.23}$$

and, within this interval, the equilibrium wage offer distribution is given by

$$F(w) = \frac{\delta+\lambda}{\lambda}\left[1 - \sqrt{\frac{p-w}{p-b}}\right] \tag{2.24}$$

The equilibrium wage distribution across workers, $G(w)$, is given by

$$G(w) = \frac{\delta}{\lambda}\left[\sqrt{\frac{p-b}{p-w}} - 1\right] \tag{2.25}$$

The expected wage, $E(w)$, is given by

$$E(w) = \frac{\delta}{\delta+\lambda}b + \frac{\lambda}{\delta+\lambda}p \tag{2.26}$$

Proof. See Appendix 2.

Let us now discuss some implications of these results.

2.5 Perfect Competition and Monopsony

The formulae for the equilibrium wage offer distribution and the wage

distribution are not very intuitive.[8] But the formula for the expected wage is simple, saying that the expected wage is a weighted average of the marginal product of labor and the reservation wage (the value of leisure), the weight on the marginal product being an increasing function of (λ/δ), the ratio of the arrival rate of job offers to the job destruction rate.

In equilibrium, all workers get paid a wage below their marginal product (note that the upper bound for wages in (2.23) is below p). This contrasts with the perfectly competitive labor market in which workers receive a wage equal to their marginal product. One might wonder about the relationship between the equilibrium here and the perfectly competitive equilibrium. It turns out that perfect competition is a special case in which job offers arrive infinitely fast for employed workers.

Proposition 2.6. *As $(\lambda/\delta) \to \infty$, the distribution of wages across workers collapses to the perfectly competitive equilibrium in which all workers get paid their marginal product, p.*

Proof. Take the limit of (2.26) and note that $E(w) = p$ implies all workers get paid p as no workers ever get paid more than p.

This proposition corresponds well with our notion of perfect competition as a market in which there is fierce competition among employers for workers and the high arrival rate of job offers means that the threat of workers leaving if they are paid a low wage is a very real one. As the result implies that perfect competition is a special (but extreme) case of our labor market, conclusions reached using a competitive analysis are not inevitably wrong; they will be correct or nearly correct if labor market frictions are small. But it is important to correct the impression that those who believe that employers have some market power over workers are extremists—the reality is that those who believe in perfect competition are the fanatics as perfect competition is one point at the edge of the parameter space and every other point in the parameter space gives employers some monopsony power. But, although it is extreme to assume the labor market is frictionless, it may be that this is a good approximation to reality if the frictions are "low." It would be helpful to have some quick way of deciding the extent of monopsony power possessed by employers.

[8] Indeed, they should not be taken too literally as the wage distribution of (2.25) has an increasing density, a prediction that is at variance with empirical observation. There is a literature (see van den Berg and Ridder 1998; Mortensen 1998; Bontemps et al. 1999, 2000) which extends the basic model to make its predictions more consistent with the observation while preserving its qualitative features. This is discussed in more detail in section 4.8.

2.6 A Simple Measure of Monopsony Power

What limits the ability of employers to lower wages is the ability of workers to leave for another employer. So, one way to understand the result in Proposition 2.6 is that a high arrival rate of job offers makes the workers' quit threat more powerful and increases direct competition among employers for workers. The extent to which workers do freely move among employers is then likely to be a good way to measure the extent of competition in the labor market.[9] But the separation rate itself is not a good measure of labor market competition as it does not matter much to employers if workers quit freely if there is a high flow of workers recruited from non-employment to replace them. In terms of Proposition 2.6, one can see that it is (λ/δ) that is important and not just λ. A simple statistic that captures this idea is the proportion of recruits that come from other firms. The higher this proportion the more intense the competition among employers and the lower we would expect the extent of monopsony to be.

This is likely to be a good "back-of-the-envelope" measure of the extent of labor market competition in many models of the labor market but the following proposition verifies the intuition by showing that, in the simple Burdett–Mortensen model, the proportion of recruits from non-employment is a monotonic function of (λ/δ).

Proposition 2.7. *The higher the fraction of recruits from non-employment, the more monopsonistic is the labor market. The fraction of recruits from non-employment in the Burdett–Mortensen model is given by*

$$\frac{\lambda}{\delta+\lambda}\frac{1}{\ln\left(\frac{\delta+\lambda}{\delta}\right)} \tag{2.27}$$

and is monotonically decreasing in (λ/δ).

Proof. See Appendix 2.

(2.27) demonstrates that the fraction of recruits from non-employment is a function of the ratio of the job offer arrival rate to the job

[9] This implicitly assumes that threats to quit if paid low wages do, in equilibrium, turn into actual quits. Some economists are inclined to argue that threats can be important even if they are never actually carried out so may not like the statistic proposed here to measure the extent of competition in labor markets. But, Proposition 2.7 shows that, although the limiting competitive case of the Burdett–Mortensen model has no wage dispersion among workers, there is a very large amount of actual worker mobility that lies behind this.

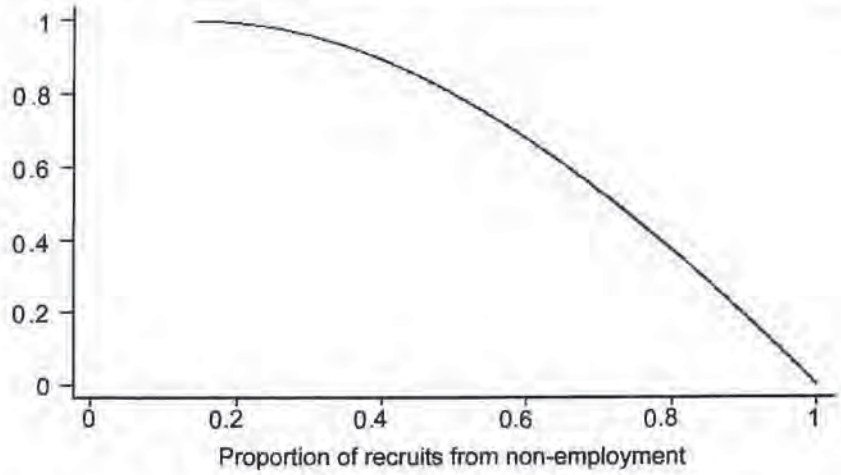


Figure 2.2 The relationship between the weight on marginal product and the proportion of recruits from non-employment.
Notes. The weight on the marginal product in the expected wage is derived from (2.26) so the left-hand axis is $\lambda/(\lambda + \delta)$. The relationship between this variable and the fraction of recruits from non-employment is given by (2.27).

offer destruction rate so is related to the indices of monopsony power described above. The relationship between the fraction of recruits from non-employment and $\lambda/(\lambda + \delta)$ is represented in figure 2.2. Remember that, from (2.26), $\lambda/(\lambda + \delta)$ is also the weight on the marginal product in the expression for the expected wage and the labeling of figure 2.2 reflects this. As one can see, the relationship is non-linear: if the proportion of recruits from non-employment is below 25% the weight on productivity in the expected wage will be above 98%. But if the proportion of recruits from non-employment is 50%, the weight will be only 80%. This discussion suggests that a simple but crude way of getting some a priori idea of the extent of monopsony is to examine the proportion of recruits that come from non-employment.

The main labor market surveys, the Current Population Survey (CPS) for the United States and the Labour Force Survey (LFS) for the United Kingdom, can be used for this purpose. Both the CPS and LFS are rolling panels: in the CPS, individuals are in the sample for four consecutive months, followed by four months out and then another four months in. In the LFS, individuals are in the sample for five successive quarters. Neither the CPS nor the LFS allow us to directly observe the fraction of recruits that were previously in non-employment as they do not contain continuous data on employment. Our approach is to approximate the proportion by considering new recruits and recording whether their labor market status

at the previous wave was employment or non-employment. As the time interval between observations on labor market status in the CPS is only a month, this is likely to be a reasonably good estimate.[10] The quarterly time interval in the LFS is perhaps more problematic.

Both the CPS and the LFS are address-based surveys so that individuals who move address between surveys are dropped from the sample. This is not a problem if the fraction of recruits from non-employment is the same for those who change address and those who do not, but one would like to be reassured on this. Fortunately, the LFS does contain information that allows us to infer labor market transition rates for those who move address. We would expect that, for every individual who leaves a sample address (a "mover-out" in the jargon), there is someone who moves into a sample address (a "mover-in"). As all employed workers are asked about their job tenure in their current job, and those who report being in their current address less than 3 months are asked about their labor market status 3 months ago,[11] one can use this information to compute the fraction of recruits from non-employment for the movers-in. In practice, this makes very little difference (47% of new recruits who are movers-in having come from non-employment as compared to 44% for the residential stayers).

Table 2.1 presents some statistics on the fraction of recruits from non-employment using the data sets described above. In the CPS, the fraction of new recruits who were not employed a month ago is 55%. This proportion includes the 5% who reported being on temporary lay-off last month. Quite how those on temporary lay-offs should be treated is

[10] In fact, in a labor market in steady-state one can show that the observed proportion of hires from non-employment in a period of unit length is given by $\xi[1 - \exp(-(\delta/\xi)(\xi + (1 - \xi)\theta))]/[1 - \exp(-(\delta/\xi))][(\xi + (1 - \xi)\theta)]$ where ξ is the true proportion, θ is the ratio of the rate at which the non-employed find jobs to the rate at which the employed change jobs, and δ is the rate of entry into non-employment. There is a bias to the extent that θ differs from 1. As one would expect that $\theta > 1$, one can show that one is likely to understate the proportion coming from non-employment.

[11] There is some reason to believe that the responses to such retrospective questions understate labor market transition rates. For the LFS we can get some information on this as, each spring, individuals are asked about their labor market status one year ago. For those in the final wave, this answer can be compared to the one they gave a year ago in the first wave. Overall the accuracy is high: over 95% of individuals for whom we have panel information and retrospective information gave answers to the retrospective questions that were consistent with the panel information. But this overall figure hides an important difference in the consistency of response: for those whose labor market states were the same in the two years (assuming the panel information is correct) about 98% gave consistent answers. But, for those who had made a labor market transition, the proportion of consistent answers fell to about 78% (the nature of the transition does not matter). However, we are using information on quarterly transitions here so we might expect this problem to be less serious.

TABLE 2.1
The Proportion of Recruits from Non-employment

Country	*Data Set*	*Sample*	*Period of Observation of Labor Market Status*	*Fraction of Recruits from Non-employment*
US	CPS	1998–99, age 18–60	Monthly	0.551
US	CPS	1998–99 (ignoring temp lay-offs), age 18–60	Monthly	0.525
UK	LFS	1992–99, age 18–60	Quarterly	0.443
UK	BHPS	1990–98, age 18–60	Continuous	0.465

Notes.
1. The fraction of recruits from non-employment is computed by taking all those in new jobs and computing the fraction for whom the economic activity before starting the job was non-employment.

not clear and it is hard to work out whether those now in employment have returned to the job from which they were originally laid off. So, the second row simply excludes them: this slightly lowers the proportion previously non-employed to 52%. The third row reports UK estimates from the LFS: the fraction of recruits from non-employment is lower than in the United States, approximately 45%. One might be concerned that this is the result of the fact that observations on employment are only at quarterly intervals. However, the fourth row of table 2.1 reports estimates from the British Household Panel Survey (BHPS) which has a continuous record of employment. The estimate of the proportion of recruits from non-employment is reassuringly similar to that derived from the LFS. So, it does seem that a lower proportion of recruits come from non-employment in the United Kingdom as compared to the United States: this is perhaps the result of the lower flows from employment into non-employment in the United Kingdom.

The aggregate figures in table 2.1 hide a lot of variation across individuals and over time. This is investigated in table 2.2 where probit models for whether an individual has been recruited from non-employment are reported. We first estimate equations for men and women jointly and then separately as there are important differences.

For the United States, table 2.2 suggests that the fraction of recruits from non-employment is higher for women, for young and very old workers, for the less-qualified, for those in full-time education, and for blacks (although the difference is only significant for men). Table 2.2 also shows

TABLE 2.2
Disaggregated Analysis of the Proportion of Recruits from Non-employment

	(1) US: CPS All	(2) US: CPS Women	(3) US: CPS Men	(4) UK: LFS All	(5) UK: LFS Women	(6) UK: LFS Men
Female	0.099			0.037		
	(0.002)			(0.003)		
Experience 1–5 years	0.036	−0.007	0.080	−0.057	−0.091	−0.029
	(0.004)	(0.009)	(0.010)	(0.005)	(0.007)	
Experience 6–10 years	−0.015	−0.024	−0.023	−0.062	−0.039	−0.087
	(0.004)	(0.010)	(0.011)	(0.006)	(0.008)	
Experience 11–20 years	−0.003	0.025	−0.032	−0.045	0.012	−0.111
	(0.004)	(0.008)	(0.010)	(0.005)	(0.007)	
Experience 31–40 years	0.024	0.010	0.046	−0.039	−0.038	−0.040
	(0.004)	(0.010)	(0.012)	(0.006)	(0.009)	(0.009)
Experience 41+ years	0.066	0.062	0.049	0.098	0.127	0.077
	(0.011)	(0.027)	(0.027)	(0.009)	(0.014)	(0.012)
High school dropout (US), no qualifications (UK)	0.103	0.109	0.108	0.068	0.074	0.065
	(0.004)	(0.009)	(0.009)	(0.005)	(0.007)	
Some college (US), A levels (UK)	−0.046	−0.064	−0.035	−0.051	−0.026	−0.060
	(0.003)	(0.008)	(0.008)	(0.005)	(0.007)	(0.006)
College degree	−0.115	−0.102	−0.108	−0.059	−0.059	−0.069
	(0.004)	(0.008)	(0.009)	(0.005)	(0.005)	(0.007)
Student	0.158	0.065	0.235	0.244	0.227	0.270
	(0.005)	(0.012)	(0.011)	(0.006)	(0.008)	(0.009)
Black	0.033	0.020	0.103	0.106	0.101	0.112
	(0.018)	(0.039)	(0.038)	(0.014)	(0.019)	(0.021)
Hispanic (US), Asian (UK)	−0.056	−0.031	−0.048	0.104	0.104	0.103
	(0.018)	(0.038)	(0.038)	(0.011)	(0.016)	(0.015)
Male employment/ population ratio	−0.130	−0.132	−0.105	−0.093	−0.015	−0.233
	(0.082)	(0.112)	(0.118)	(0.153)	(0.215)	(0.219)
Observations	172464	88797	83667	96086	48644	47442
Mean of dependent variable	0.545	0.587	0.501	0.443	0.463	0.422
Pseudo-R^2	0.032	0.021	0.041	0.040	0.035	0.057

Notes.

1. The sample is all those who have just started jobs. The dependent variable is a binary variable taking the value one if the individual was recruited from non-employment.
2. The sample period is 1994–2000 for the CPS and 1992–2000 for the LFS.
3. Reported coefficients are marginal effects. Standard errors in parentheses.
4. Regressors are, as far as possible, defined in the same way for the US and UK data. Where there are unavoidable differences (in education and race), the appropriate variables are defined in the first column.
5. Student is defined as anyone who has not completed full-time education. Qualifications are coded as zero for these individuals. The omitted education category for the United States is a high school graduate and for the United Kingdom someone with O levels.
6. The CPS regressions also include month, year and state dummies. The LFS regressions also include month, year and region dummies.

very similar results for the United Kingdom. Women also have a rather different experience profile in both countries, being more likely to have been recruited from non-employment when they have between 11 and 20 years of experience: this is likely to be associated with withdrawal from and return to the labor market connected with having children. What is worth noting is that those groups that do badly in the labor market in terms of wages also do badly in terms of more frequently being recruited from non-employment. This is in line with the basic prediction of the theory where competition among employers for workers is more intense (and wages end up higher) when a lower fraction of recruits are from non-employment.

The estimated models in table 2.2 also include the prime-age male employment/population ratio to see whether there is cyclical variation in the proportion of recruits from non-employment. For the United Kingdom, the results in Burgess (1993) suggest that the proportion of recruits from non-employment falls as the labor market tightens as job-to-job moves are very pro-cyclical (although Fallick and Fleischman 2001 conclude this is not true for the United States). The results in table 2.2 provide some weak support for this conclusion although the coefficient is never significantly different from zero.[12]

In this section we have proposed that the fraction of recruits from non-employment is a crude but simple measure of the extent of competition among employers for workers. Differences in this measure across different types of workers also mirror wage differences as the theory predicts (although others might also do so).

2.7 Positive and Normative Aspects of Monopsony and Oligopsony

In an oligopsonistic equilibrium, workers are "exploited" in the sense of that term used by Hicks and Pigou: that is, they receive a wage less than their marginal product. But, the word "exploitation" has emotive power that is unfortunate in the current context. In the static model of monopsony it makes sense to use the marginal product of labor as a point of comparison for the wage. The efficient outcome is to set the wage equal to the marginal product and a minimum wage set at that level leads to a first-

[12] Although one explanation for this is that most of the variation in the employment/population ratio is absorbed by the time and regional dummies. The United Kingdom does show a remarkable fall in the proportion of recruitments from non-employment from almost 50% in 1992 to just above 35% in 1999, consistent with the rise in employment in the same period. However, there is no noticeable trend in the fraction of recruits from non-employment in the United States over a similar period.

best outcome. However, that is not necessarily true in the models of oligopsony discussed here. Even though Proposition 2.7 says that, in a frictionless market, workers would get paid their marginal product, one cannot wish away the existence of frictions and, given their existence, it is not clear that efficiency would be best served by raising wages to the marginal product. The bulk of this book is positive: about how we can achieve a better understanding of a wide variety of labor market phenomena from the distribution of wages to the provision of training to the impact of minimum wages and trade unions by recognizing that employers have non-negligible market power over their workers. There is little normative content: no judgment is made about whether these things are "good" or "bad," although, as we shall see, the approach taken here does suggest approaching many issues with a more open mind than a fanatical believer in perfect competition might be inclined to do. This emphasis on the positive aspects of the subject is not because the normative issues are unimportant but because the normative concerns are sufficiently complex that it is simply not credible to be able to draw normative conclusions from theoretical introspection or from casual empirical analysis. Justifying this conclusion is the subject of the next chapter: this can be skipped for those who are only interested in the positive implications of oligopsony.

2.8 Implications and Conclusions

The models of this chapter have been highly stylized with assumptions chosen for analytical convenience more than for realism. Nonetheless, they do convey the essence of a labor market with frictions in which employers set wages influenced in part by competition from other employers but in which this competition is not so cutthroat as to enable workers to extract all the surplus from the employment relationship, nor so feeble as to enable employers to get all the surplus.

A lot of the analysis in this chapter has been very formal. But, one should not allow this to distract attention away from the basic insights into the workings of labor markets that the monopsonistic approach provides. The rest of this book is concerned with the determinants of prices and quantities in the labor market, a traditional pre-occupation of microeconomics. The study of prices is essentially the study of the distribution of wages while the study of quantities is the study of the level and distribution of unemployment, the level of employment in firms, and of the quality of labor (as influenced by the acquisition of human capital). The implications of monopsony or oligopsony are summarized briefly for these issues.

In perfect competition we normally think of the distribution of wages as being determined by the distribution of marginal products[13] and attempt to explain wage differentials in cross-section and over time in this way. In a monopsonistic labor market, marginal productivity continues to be an important explanation of wages but other factors are also important. Perhaps the simplest way to see this is to look at the expression for the expected wage in (2.26). Marginal product, p, appears but so does:

- the value of leisure (the reservation wage);
- job offer arrival rates;
- job destruction rates.

As monopsony gives the labor economist a wider menu of possible explanations of the distribution of wages, we might hope for a richer explanation than can be provided when constrained by the straitjacket of perfect competition. In chapters 5 through 8, we show how such an approach can improve our understanding of the distribution of wages. For example, chapter 7 attempts to explain part of the gender wage gap in terms of the different labor market transition rates of men and women. There is also one final factor that is important in influencing wages: luck. The existence of wage dispersion among identical workers means that there is likely to be some part of the distribution of wages that can never be explained by economic factors: some workers will simply have been in the right place at the right time.

The differences in the determinants of quantities in the labor market might appear to be less dramatic, largely because search models are already commonly used to understand both theoretical and empirical aspects of unemployment. For example, in the model presented in this chapter, the level of non-employment is influenced by the job offer arrival rate when non-employed, the job destruction rate, and the level of wages relative to the reservation wage. This is not very different from the usual list of suspects although a strict perfect competition approach would suggest that only a comparison of the marginal product with the value of leisure is relevant.

The rest of the book aims to demonstrate how we can gain a better understanding of labor markets by this less dogmatic approach based on the perspective that employers have some market power.

Before we move on to these positive and empirical issues, the next chapter is concerned with more normative concerns, for example, is the oligopsonistic labor market efficient? and, if not, is the inefficiency of any

[13] Abstracting from compensating wage differentials (discussed in chapter 8) and the fact that marginal products may themselves be endogenous, varying with the level of employment.

particular type? and are there any policy interventions that might be expected to improve the operation of the labor market? This discussion is entirely theoretical: for those uninterested in it, one can summarize the conclusions now:

- the oligopsonistic labor market is not generally efficient;
- it is hard a priori to say anything about the direction of the inefficiency;
- it is hard to make a strong theoretical case for any particular policy intervention.

These conclusions then justify the approach in the rest of the book which is to use the perspective of employer market power to understand a wide variety of labor market phenomena, without making any value judgment as to whether the world could be improved by an appropriate policy intervention.

Appendix 2

Proof of Proposition 2.1

At any date t the state variable for the firm will be the labor force that it had last period, N_{t-1}. Define a value function $\Pi(N_{t-1})$ to be the maximized discounted value of future profits from date t onwards. So, using dynamic programming arguments, we have

$$\Pi(N_{t-1}) = \max_{(w_t, N_t)} Y(N_t) - w_t N_t + D\Pi(N_t) \tag{2.28}$$

This needs to be maximized subject to a dynamic labor supply curve of (2.4). Note that this dynamic labor supply curve depends only on the current wage: this implicitly assumes workers are myopic. An alternative would assume it depends on the value of the job.

Taking the first-order condition of (2.28) with respect to w_t and taking account of the dependence of N_t on w_t leads to the following first-order condition:

$$[Y'(N_t) - w_t + D\Pi'(N_t)]\frac{\partial N_t}{\partial w_t} - N_t = 0 \tag{2.29}$$

We also have the envelope condition which allows us to derive the derivative of the value function. Differentiating (2.35) we have that

$$\Pi'(N_{t-1}) = [Y'(N_t) - w_t + D\Pi'(N_t)]\frac{\partial N_t}{\partial N_{t-1}} \tag{2.30}$$

If the firm is in a steady state (and it is an interesting question whether there is a steady state) where wages and employment are constant, then

we can solve (2.30) for $\Pi'(N)$ which leads to

$$\Pi'(N) = \frac{\frac{\partial N_t}{\partial N_{t-1}}[Y'(N) - w]}{1 - D\frac{\partial N_t}{\partial N_{t-1}}} \tag{2.31}$$

Substituting this into (2.29) and re-arranging, one can derive

$$\begin{aligned}\frac{Y'(N) - w}{w} &= \left[1 - D\frac{\partial N_t}{\partial N_{t-1}}\right]\frac{N_t}{w(\partial N_t/\partial w_t)} \\ &= \left[1 - D + D\left(1 - \frac{\partial N_t}{\partial N_{t-1}}\right)\right]\frac{N_t}{w(\partial N_t/\partial w_t)} \\ &= (1 - D)\varepsilon^s + D\varepsilon \end{aligned} \tag{2.32}$$

where the last equality follows from the fact that differentiation of (2.4) implies that

$$1 - \frac{\partial N_t}{\partial N_{t-1}} = s$$

and the relationship between the short- and long-run elasticities implied by (2.8).

Proof of Proposition 2.2

Suppose there is a mass of firms offering the wage w. If $w = p$, then all these firms must be making zero profits. A firm that lowers its wage can make higher profits as long as it can retain some workers in steady state, that is, as long as its separation rate is finite and its recruitment rate positive. A non-zero, finite value of (λ/δ) guarantees this.

If there is a mass of firms paying $w < p$ then consider what happens if a firm deviates by paying an infinitesimally higher wage. Profit per worker is only infinitesimally reduced but the number of workers is measurably higher (as long as $\lambda > 0$) as recruits now come from workers in all the firms who continue to pay w. Hence, profits must rise and the initial situation could not have been in equilibrium.

Proof of Proposition 2.3

The simplest way to prove Proposition 2.4 is by equating inflows and outflows from the group of workers earning w or less. The outflow rate from this group will be $[\delta + \lambda(1 - F(w))]$ as workers leave the group either to non-employment or to better-paying jobs. Recruits to this group must

come from non-employment as no workers who earn more than w will ever accept a wage offer less than w. There are $M_w u$ non-employed workers who receive offers less than w at a rate $\lambda F(w)$. So the flow of recruits to jobs paying less than w will be $\lambda F(w) M_w u$. Equating inflows and outflows, we then have

$$[\delta + \lambda(1 - F(w))](1 - u)G(w; F)M_w = \lambda F(w)uM_w \qquad (2.33)$$

as total employment of those earning w or less will be $(1 - u)G(w; F)M_w$. Using (2.17) leads, after some re-arrangement, to (2.18).

Proof of Proposition 2.4

Suppose a firm pays below b. This firm will have no workers so will make zero profits which cannot be an equilibrium.

Suppose the lowest wage offered is strictly above b. The lowest-wage firm will only recruit workers from non-employment at a rate $(\lambda u/M)$ and will lose workers whenever they get another job offer, that is, at a rate $(\delta + \lambda)$. So, employment in the lowest-wage firm will be given by

$$\frac{\lambda u}{M[\delta + \lambda]} = \frac{\delta\lambda}{M[\delta + \lambda]^2} \qquad (2.34)$$

that is, independent of the wage offered. If the lowest-wage firm cuts its wage (but not below b) the recruitment and separation rate will be unchanged and hence so will employment. But, profit-per-worker will rise so profits will increase. This means the original situation could not have been in equilibrium.

Proof of Proposition 2.5

Equating (2.21) and (2.22) leads to (2.24) after some re-arrangement. The right-hand side of (2.23) is then just the value of the wage that makes $F = 1$. (2.25) comes from (2.18) and (2.24).

Now the expected wage can be written as

$$E(w) = \frac{M_f}{M_f}\frac{\int wN(w; F)f(w)dw}{\int N(w; F)f(w)dw} = p - \frac{M_f \int (p - w)N(w; F)f(w)dw}{M_w(1 - u)}$$

$$= p - \frac{M \int \pi^* f(w)dw}{(1 - u)} = p - \frac{M\pi^*}{(1 - u)} = p - \frac{\delta(p - b)}{\delta + \lambda} \qquad (2.35)$$

where the second equality follows from the fact that the two denominators in the first line are both expressions for total employment; the third equality follows from the fact that, in equilibrium, $(p - w)N$ must be the same for all firms; and the final equality follows from (2.22) and (2.17).

Proof of Proposition 2.6

The recruits to position F in the wage distribution, $R(F)$, can be written as

$$R(F) = \lambda u M_w + \lambda M_w(1-u)G(F) = \frac{\lambda \delta M_w}{\delta + \lambda(1-F)} \tag{2.36}$$

where $G(F)$ is the fraction of workers employed at position F or below and is given by (2.18). Note that the position in the wage distribution is a sufficient statistic for the number of recruits so we do not have to worry about the actual wage paid. Using the fact that F must be distributed uniformly over the unit interval, we have that the total flow of recruits in the economy is given by

$$R = \int_0^1 R(f)df = M_w \int_0^1 \frac{\lambda \delta df}{\delta + \lambda(1-f)} = M_w \delta \ln\left(\frac{\delta + \lambda}{\delta}\right) \tag{2.37}$$

As the flow of recruits from non-employment is $\lambda M_w u$, this gives (2.27) for the fraction of recruits from non-employment.

3

Efficiency in Oligopsonistic Labor Markets

Discussions of the partial equilibrium static model of monopsony often emphasize that the free market equilibrium is inefficient in a very particular way. Both the wage and employment are too low and full efficiency can be restored by ensuring that the wage is equal to what it would be in a perfectly competitive labor market. One way of achieving this outcome is by means of an artfully chosen minimum wage. This chapter considers whether general equilibrium models of oligopsony allow such clear-cut policy prescriptions: the conclusion is that they do not, although there is no presumption that the "free market" equilibrium is efficient.

This chapter discusses the efficiency issue using the Burdett and Mortensen (1998) model introduced in chapter 2. The simple version of the model presented in the previous chapter cannot provide an adequate analysis of efficiency because the equilibrium is fully efficient as all matches between unemployed workers and firms are consummated. Any distribution of the surplus between employer and workers is consistent with this equilibrium outcome. For example, any minimum wage up to the level of p, the marginal product of workers, results in the same outcome in terms of employment and only affects the division of the surplus between wages and profits.

But, this conclusion is the result of simplifying assumptions made for expositional reasons. Efficiency is not an interesting issue in this version of the model because very few decisions of employers and workers are free to respond to incentives. However, there are a number of ways in which one might modify the model to endogenize decisions of both firms and workers so that they can be affected by incentives.

First, the model of the previous chapter assumed that the supply of both workers and firms to the market is inelastic, that is, the number of workers and firms in the market are fixed at M_w and M_f, respectively. A natural way to introduce incentives is to assume that the supply of both firms and workers is not inelastic. For firms, the simplest way to do this is to assume that there is free entry (i.e., to go to the opposite extreme and to assume that the supply of firms to the market is perfectly elastic). For workers, one could make an analogous assumption that a fixed cost (perhaps the cost of acquiring skills necessary for employment) must be paid to enter

the labor market but there are alternative ways to make the overall labor supply have some elasticity. For example, section 3.4 discusses the case where there is heterogeneity in the value of leisure (hence the reservation wage) so that not all workers will be interested in every job.

But, even once agents have decided to participate in the market there are other decisions that might be influenced by incentives. For example, firms can decide how much effort to spend in looking for recruits and workers can decide how hard to look for work. Jointly, these decisions about search intensity can be expected to determine the arrival rate of job offers, λ.

Agents may make some investments in match quality before a match is realized. For example, workers may make decisions about how much human capital to acquire before they start looking for a job. And firms may have to commit capital to jobs before they start looking for workers for those jobs. Both of these decisions will be affected by expected future returns to the agents.

All of these "margins" of decision are likely to be present in reality. But, to include all of them simultaneously in a model of the labor market is a recipe for indigestion. Consequently, this chapter presents only the simplest possible models to make the relevant points. The models examined in sections 3.1–3.5, their main conclusions, and their implications for one particular policy intervention (the minimum wage) are summarized in table 3.1.

TABLE 3.1
The Structure of the Chapter

Section	*Model*	*Efficiency of Free Market*	*Optimal Minimum Wage*
3.1	Free entry of firms (perfectly elastic supply of firms to the market)	Too many firms, employment too high	Minimum wage to ensure appropriate division of surplus
3.2	Endogenous recruitment activity of firms	Too much recruitment, employment too high	Minimum wage to ensure appropriate division of surplus
3.3	Free entry of workers (perfectly elastic supply of workers to the market)	Too few workers in labor force, employment too low	Minimum wage to ensure appropriate division of surplus
3.4	Heterogeneity in reservation wages	Employment too low	Minimum wage equal to marginal product
3.5	Heterogeneity in reservation wages + free entry of firms	Employment may be too high or too low	Minimum wage may not be desirable

The main conclusions of sections 3.1–3.5 are as follows:

- there are good reasons to believe the free market equilibrium is inefficient;
- it is hard to say anything a priori about the direction of the inefficiency;
- it is hard to make unambiguous predictions about policy from theoretical models alone.

Thus, issues of efficiency are more complicated and subtle in general equilibrium dynamic models of oligopsony than the static partial equilibrium textbook model of monopsony might suggest. The conclusion that theory provides little in the way of a guiding principle is, in many ways, an unsatisfying one as it suggests a failure to find the most general result. One might wonder whether one can find better guidance elsewhere in the literature.

Other "search" models of the labor market have discussed whether the free market equilibrium is likely to be efficient. This debate probably started with Friedman's (1968) celebrated article on the natural rate of unemployment. Commentators like Tobin (1972) argued that the free market equilibrium was likely to be inefficient and there have been papers on the subject ever since (e.g., Albrecht and Jovanovic 1986; Hosios 1990; Moen 1997; Acemoglu and Shimer 2000; among others). Many (though not all) of the papers appear to arrive at sweeping conclusions but readers must recognize that they are based on very specific models and are not robust to reasonable changes in the assumptions.[1] Perhaps there is a result to be derived about the efficiency or otherwise of the free market in a class of reasonably general models. Armed with such a result, there might also be general prescriptions about policies to get to the first-best (the Coase Theorem and the prescription to define property rights springs to mind as an example from another part of economics). But, a combination of lack of intellect and laziness on the part of this author means this book does not provide such a general result and it certainly does not exist in the current literature. My conjecture is that, whenever a model of the labor market is proposed that has strong conclusions, one will be able to provide another, observationally equivalent, model with different conclusions. If this conjecture is correct, theory alone is not going to help us very much. This conclusion is similar in spirit to that of Lucas and Prescott (1974, p. 206) who criticized Tobin's (1972) conclusion that the free market *must* be inefficient while recognizing that "the question of

[1] For example, Albrecht and Jovanovic (1986, p. 1256), conclude that "in contrast to the competitive equilibrium, the monopsonistic equilibrium is shown to be inefficient involving too much search and too little employment." This is, of course, a correct conclusion in their model but the model is not a general one and the conclusion would be relatively simple to overturn.

whether there exist important external effects in *actual* labour markets, remains, of course to be settled."

The plan of the chapter is as follows. Section 3.1 considers the welfare properties of the equilibrium when there is free entry of firms, section 3.2 when employers can choose their level of recruitment activity. Section 3.3 then considers the case where there is free entry of workers and section 3.4 assumes some heterogeneity in the reservation wages of workers. Section 3.5 puts together the model of the sections 3.1 and 3.4, illustrating how the whole is rather different from the sum of the parts. Section 3.6 shows how easy it is to generate multiple equilibria in models of oligopsony, a feature that is potentially useful in explaining a range of phenomena from agglomeration to ghettoes.

3.1 Free Entry of Firms

In this section the basic Burdett–Mortensen model of section 2.4 is extended to the case where there is a perfectly elastic supply of firms to the market so that the number of firms is M_f is endogenous. To produce an equilibrium with a finite number of firms one needs to assume that there is a fixed cost of entry which we will denote by C_f. Firms enter until profits are equal to C_f. Using results derived in the previous chapter, equilibrium profits are given by (2.22). The number of firms affects profits as it affects M $(= M_f/M_m)$. But it also plausibly affects λ, the arrival rate of job offers for workers. Suppose that the number of matches between workers and firms is given by $m(M_w, M_f)$. It is conventional and convenient (for a recent survey, see Petrongolo and Pissarides 2001) to assume that the matching function has constant returns to scale so that the arrival rate of job offers for workers can be written as

$$\lambda = \frac{m(M_w, M_f)}{M_w} = m(1, M) \equiv \lambda(M) \tag{3.1}$$

where $\lambda(M)$ is an increasing concave function of its argument so that a fall in the number of firms reduces the arrival rate of job offers for workers. M will be given by the level at which profits are equal to C_f. Taking account of (3.1) and (2.22), one can write this free entry condition as

$$\frac{\delta\lambda(M)(p-b)}{M[\delta + \lambda(M)]^2} = C_f \tag{3.2}$$

By differentiating the left-hand side of (3.2) one can readily verify that profits are a strictly decreasing function of M if $\lambda(M)$ is an increasing concave function of M so that (3.2) defines a unique equilibrium.

Now, let us consider the efficient level of M. If the non-employment rate is u, then a fraction $(1-u)$ of workers are in employment producing p, while a fraction u are not in work which has a value b. In addition, the total fixed costs paid in the economy are $M_f C_f$. Hence, the total social surplus can be written as

$$\Omega(M) = M_w[(1-u)p + ub] - M_f C_f = M_w\left[p - \frac{\delta(p-b)}{[\delta + \lambda(M)]} - MC_f\right] \tag{3.3}$$

where the second equality follows from the non-employment rate of (2.17).[2]

The following proposition summarizes the efficiency of the free market equilibrium.

Proposition 3.1

1. *The free market has too many firms if $\lambda(M)$ is a strictly concave function of M.*

2. *The first-best can be attained by setting a minimum wage, w_m, which satisfies*

$$\frac{p - w_m}{p - b} = \varepsilon_{\lambda M} \equiv \frac{M\lambda'(M)}{\lambda(M)} \tag{3.4}$$

where M is the efficient number of firms relative to workers.

Proof. See Appendix 3.

The intuition as to why the free market equilibrium has too many firms is that some of the employment of new entrants comes not just from workers who would otherwise be unemployed but also from those employed in other firms. While this source of employment is a private gain, it has no social purpose. The second part of the proposition shows that a well-chosen minimum wage can attain the first-best, although the case for a minimum wage is that the minimum wage causes exit of firms from the market and reduces employment but these are "good" things. This argument for minimum wages is slightly curious as proponents of minimum wages do not often argue for it on the grounds that it destroys

[2] Note that this specification of the welfare function assumes risk neutrality on the part of workers. If workers are risk averse, then the wage dispersion that is characteristic of the free market equilibrium has a welfare cost and there would be a case for policies to reduce this dispersion.

jobs. One should note that policies other than the minimum wage could be used to attain the first-best. For example, one could pay unemployment benefits to ensure that the reservation wage of workers is w_m or one could use a profits tax.

One interpretation of (3.4) is that the share of the total surplus going to the employer in the lowest-wage firm should be equal to the elasticity of the matching function with respect to the number of firms. This is outwardly very similar to the efficiency rule derived by Hosios (1990) in the context of a Diamond–Pissarides matching model in which wages are determined by an ex post sharing rule. There is one important difference, namely that (3.4) refers only to the sharing of the surplus in the lowest-wage firm. The employer's share of the surplus in all the other firms will be lower than that given in (3.4) as they all pay wages higher than w_m. This is important, because knowledge of the Hosios rule might have led one to conclude that, because workers do get a share of the surplus in the basic Burdett–Mortensen model, it is not obvious a priori whether wages are too high or too low. As the above discussion has made clear, wages are unambiguously too low.

3.2 Endogenous Recruitment Activity

In this section, firms are allowed to influence the flow of recruits by expenditure on recruitment activity. For the moment, assume that the number of firms, M_f, is exogenously given. Denote by z the intensity of the recruitment activity of the firm. Define z so that, other things being equal, the arrival rate of workers to the firm is proportional to z. Assume that the cost of z is given by a function $c(z)$. One could interpret this cost function narrowly as the cost of advertisement but it is probably better to think of it more widely as the cost of recruitment and training new employees as the administrative costs of handling applications and the induction of new workers are typically much larger than the direct costs of job advertisements.

In equilibrium there will be some function $z(w)$ which relates recruitment activity to the wage paid. Denote by Z the average level of recruitment activity in the economy as a whole, that is, assume Z is given by

$$Z = \int z(w)dF(w) \tag{3.5}$$

Make the simplifying assumption that the rate at which job offers arrive to workers depends on Z as well as M so can be written as $\lambda(Z, M)$. Assume that an individual employer's share of these matches is given by (z/Z). The equilibrium can be described by the pair of functions

$\{z(w), F(w)\}$ which give the distribution of wages across firms and the recruitment intensity associated with each wage.

The following proposition summarizes the nature of equilibrium and its efficiency.

Proposition 3.2

1. *In equilibrium, all firms recruit at the same intensity, z, which is given by*

$$\frac{\delta\lambda(z, M)(p - b)}{M[\delta + \lambda(z, M)]^2} = zc'(z) \tag{3.6}$$

2. *The free market has too much recruitment activity, if $\lambda(z, M)$ is a strictly concave function of z.*

3. *The first-best can be attained by setting a minimum wage, w_m, which satisfies*

$$\frac{p - w_m}{p - b} = \varepsilon_{\lambda z} \equiv \frac{z\lambda_z(z, M)}{\lambda(z, M)} \tag{3.7}$$

Proof. See Appendix 3.

The first part of this proposition says that (whatever their chosen wage) all firms spend the same amount on recruitment activity in equilibrium. Paying a higher wage encourages recruitment expenditure as a higher fraction of workers contacted will be interested in the job and the expected job duration of a recruit is longer. But, the profit to be made from each recruit per period is less. So, there are off-setting effects of the wage on the incentives to recruit and the proposition simply says that the equilibrium wage distribution for the case analyzed here is such that these different effects cancel out and the incentives to recruit are independent of the wage.[3]

As in the free entry case of the previous section, the intuition for the excess recruitment activity of the second part of the proposition is that the employment of an extra firm comes not just from workers who would otherwise be unemployed but also at the expense of other firms. And, as in the case of free entry, there are a number of policies that might be used to correct this inefficiency. The final part of the proposition says that a minimum wage that ensures a particular division of the surplus in the lowest-wage firm can attain full efficiency. As long as $\lambda(Z, M)$ is a strictly

[3] This intuition also suggests that the independence result will fail if there are decreasing returns to labor or if there is firm heterogeneity. See Mortensen (1998) for an analysis of this case.

concave function of z, it is optimal to have $w_m > b$. But, as in the free entry case, a binding minimum wage will reduce employment as it reduces the recruitment activities of firms. (3.7) has an obvious similarity to the Hosios rule derived for the free entry case in (3.4) except that it is now the elasticity of the arrival rate of job offers with respect to the recruitment intensity that is important. This might make us wonder what happens if we combine free entry and endogenous recruitment activity. It is fairly simple to show that a simple minimum wage now can only attain the first-best if $\varepsilon_{\lambda M} = \varepsilon_{\lambda z}$ (i.e., if the matching function can be written as $\lambda(zM)$). In this case, the two Hosios conditions (3.4) and (3.7) are identical.

So far, we have only considered decisions of firms that respond to incentives. Even then, there is reason to believe that wages are "too low" although the other side of this coin is that employment is "too high." But, as discussed in the introduction, some decisions of workers are also likely to respond to incentives. The following sections introduce this topic.

3.3 Elasticity in Labor Supply: Free Entry of Workers

The basic Burdett–Mortensen model of section 2.4 assumed that the supply of workers to the labor market is inelastic. This section introduces some elasticity into labor supply in a very simple way: by assuming that, to participate in the labor market, individuals must pay an up-front cost of C_w. This is a crude way of introducing some elasticity in labor supply (an alternative is discussed in the next section), but it does have the virtue of being a natural analogy to the way some elasticity in the supply of firms to the market was introduced in section 3.1. If pushed, one could interpret the fixed cost as the cost of acquiring the human capital necessary to get employment or as an investment in the skills necessary for job search.

Let us distinguish between the value of being unemployed V^u, the value of non-participation V^n, and the value of being employed at wage w, $V(w)$. The value functions are given by

$$\delta_r V^u = b + \lambda \int_{w_m} [V(x) - V^u] dF(x) \tag{3.8}$$

$$\delta_r V^n = b + C_w \tag{3.9}$$

$$\delta_r V(w) = w - \delta_u [V(w) - V^u] + \lambda \int_w [V(x) - V(w)] dF(x) \tag{3.10}$$

where w_m is the lowest wage. (3.9) captures the fact that those who choose non-participation also save the fixed cost C_w. Free entry of work-

ers means that, in equilibrium, we must have $V^u = V^n$. The following proposition summarizes the important results.

Proposition 3.3

1. *The free market equilibrium has too few workers in the market.*
2. *The first-best can be attained by setting a minimum wage such that:*

$$\frac{p - w_m}{p - b} = \varepsilon_{\lambda M} \tag{3.11}$$

Proof. See Appendix 3.

(3.11) should, by now, be a familiar formula. The share of the surplus in the lowest-wage firm should, for efficiency, be equal to the elasticity of the arrival rate of job offers with respect to M. Because the share of workers in the surplus in the lowest-wage firm is zero in the free market equilibrium, wages are in some sense "too low" and too few individuals choose to participate in the labor market.

The effect of a minimum wage on this labor market differs from that in the labor market with free entry of firms. A binding minimum wage causes more individuals to participate in the labor market and total employment rises so that the employment/population ratio rises. But, the unemployment rate also rises as the increase in the number of agents in the labor market causes some crowding-out of job opportunities.

This section has made the supply of workers to the market perfectly elastic. Without further modification, one could not combine this model with free entry of firms as the scale of activity in the economy would then be indeterminate (for versions of this model in which there is an arbitrary degree of elasticity in the supply of both firms and workers to the market, see Manning 2001b). The next section takes a different approach to introducing some elasticity into the supply of labor to the market.

3.4 Elasticity in Labor Supply: Heterogeneity in Reservation Wages

In the models discussed so far, all workers have been assumed to have the same value of leisure, b, which, given our other assumptions, is also the reservation wage. This section modifies the model and assumes that there is some heterogeneity in b. Denote the cumulative density function of b by $H(b)$ and the associated density function by $h(b)$. It is helpful (although

not essential) to assume that $H(b)$ is log-concave. For convenience, assume that $H(p) = 1$ so that all workers have $b \leq p$.[4] Furthermore, assume that b is not observed by employers so that wage offers cannot be conditional on it. As a result, not all wages that are offered in equilibrium will be attractive to all workers. If a worker has value of leisure b (which will also be their reservation wage), then only a fraction $[1 - F(b)]$ of jobs will be desired.

The following proposition analyzes this case.

Proposition 3.4

1. *The free market equilibrium is inefficient with employment too low.*

2. *A minimum wage equal to p can restore full efficiency.*

Proof. See Appendix 3.

In this model, employment is, in general, too low. It is efficient to consummate all matches with $p \geq b$ yet matches are only consummated when $w \geq b$. Because $w < p$, some efficient matches are not consummated. If the number of firms (and their recruitment activities) are fixed, then attaining the first-best is simple: set a minimum wage equal to p and the inefficiency disappears. As in the other models discussed in this chapter, there is an efficiency case for a minimum wage but, now, a binding minimum wage is associated with increases in employment. This model is the closest oligopsony model to the static monopsony model as both employment and wages are too low in the free market equilibrium and full efficiency can be restored by ensuring wages are equal to marginal products.

However, this discussion has assumed that the supply of firms is completely inelastic: the next section considers what happens when we introduce an elastic supply of firms as in the model of section 3.1.

3.5 Heterogeneity in Reservation Wages and Free Entry of Firms

In this section we combine the model of the previous section with the model of section 3.1 in which there is a completely elastic supply of firms to the market.

One might have thought that because a minimum wage can improve efficiency in the two constituent models (the free entry model of section 3.1

[4] If $b > p$ there is no point in a worker being in the labor market as they will never be able to find employment at a wage acceptable to them.

and the heterogeneous reservation model of section 3.4), it must be of potential benefit in the current model. But, the following proposition shows that this is not necessarily the case.

Proposition 3.5

1. *Social surplus may be increasing or decreasing in the number of firms.*

2. *A just-binding minimum wage may increase or reduce efficiency.*

Proof. See Appendix 3.

This result is an example of the general principle of the second-best, that moving towards the first-best in one dimension may worsen welfare. In this case, a binding minimum wage, because it reduces the number of firms, tends to reduce the social surplus if there are too few firms in equilibrium, and may reduce the social surplus. The intuition for the ambiguity about the optimal number of firms is that, on top of the congestion effect which tends to lead to too many firms in the free market equilibrium, the entry of a new firm now improves the wage offer distribution which results in more workers being in employment thus increasing efficiency.

What policies can attain efficiency in this case? It should be readily apparent that a simple minimum wage can no longer lead to the first-best outcome. To consummate all efficient matches, the minimum wage would need to be equal to p but this results in no firms entering the market. So, a minimum wage of p would need to be combined with a subsidy to the entry of firms.

It should be apparent that we have quickly arrived at a point where theory provides little guidance about the nature of inefficiency in the free market equilibrium and the types of policies that might help. One could proceed further to consider more complicated models combining all the decisions we have endogenized here and even adding new ones. But, the payoff from this strategy is likely to be small as the qualitative conclusions we have drawn are unlikely to be altered.

3.6 Multiple Equilibria in Models of Oligopsony: An Application to Ghettoes

All the oligopsony models presented so far in this chapter have a single equilibrium. But, it is important to realize that it is relatively simple to produce oligopsony models with multiple equilibria that may further complicate welfare analysis. This is because of the way in which supply

and demand factors interact in markets with frictions: it is possible that, in a sense to be made clearer below, "supply can create its own demand" (and vice versa).

In a frictionless competitive market, there is a sense in which it is good to be unique (because of diminishing marginal productivity). If one acquires some specialized skill that requires some specialized capital with which to work, there is no problem in meeting the person with that capital or in inducing someone to make that specific capital investment. However, in a market with frictions, uniqueness is not necessarily an advantage. If employers have to make some ex ante investment in creating jobs, they may not choose to make that investment if the chance of finding a suitable person to fill those jobs is very low. In that case, an increase in the supply of workers of a particular type may encourage employers to create jobs tailored for that type of worker and hence create its own demand. The problem is that there is no mechanism to ensure that an individual act of investment by either worker or employer will be matched by the equivalent investment on the other side of the market that is necessary for the investment to have its full effect. This problem is a potent source of multiple equilibria.

There are numerous examples of this type of model in the economics literature, probably beginning with Diamond (1982) but including Acemoglu (1998) and Machin and Manning (1997), among others. They can be used to explain a number of potentially important facts: why some countries are industrialized and others are not, and the phenomenon of agglomeration.

Here, we illustrate these ideas by presenting a simple model of how ghettoes may arise. The stylized picture of the ghetto is that of an area where both wages and employment are lower than in neighboring areas. This does seem to have something to do with the disadvantage experienced by certain ethnic groups as, in the United States, inclusion of family background variables plus good measures of educational attainment (e.g., the AFQT in the NLSY) seems to be able to eliminate much, if not all, of the observed black–white wage differential (see Neal and Johnson 1996; Altonji and Blank, 1999).[5] This suggests paying attention to pre-market factors more than labor market outcomes in looking for the origins of the black–white wage differential. Some of these pre-market factors (e.g., family background) may be the product of more explicit racial discrimination that undoubtedly existed in the past, while others may be the result of the poorer quality of education typically received by blacks. But, it is also possible that some of it represents decisions not to acquire human capital that are rational given the economic situation faced.

[5] Although there is a debate about whether the AFQT test scores themselves contain a racial bias.

As a simple example of this type of mechanism, consider the Burdett–Mortensen general equilibrium model of an oligopsonistic labor market discussed in section 2.4. But modify it, so that workers, before they enter the labor market, are assumed to have a choice about the level of human capital they acquire. This will determine their productivity p. We assume that acquiring productivity of p requires a cost of $c(p)$. We also assume that the distribution of wages facing a working of productivity p is $F(w;p)$ as given by (2.24). This implies that firms offer a different wage to workers of each quality level.

Given this, p will be chosen to maximize $V^u - c(p)$. In a frictionless market, workers with quality p would earn p with probability 1 for their entire life of expected length $(1/\delta_r)$. Hence, workers would invest to the point where $\delta_r c'(p) = 1$.

In a labor market with frictions we can prove the following proposition.

Proposition 3.6

1. *The optimal choice of p is given by*

$$\delta_r c'(p) = \left(\frac{\lambda}{\delta + \lambda}\right)^2 \tag{3.12}$$

assuming an interior solution.

2. *The optimal p is increasing in λ tending to the competitive level as $\lambda \to \infty$.*

Proof. See Appendix 3.

Unsurprisingly, Proposition 3.6 says that the faster the rate of job offers arrive, the greater the incentive to invest in human capital and that human capital investment will approach the competitive level as the labor market becomes frictionless. So we would expect p as a function of λ to be something like the line marked "investment decision of workers" in figure 3.1.

Now consider the investment level by firms. Let us model this investment decision as a simple entry decision. Assume, as in section 3.1, that entry requires a fixed investment of cost C_f, and that M_f firms enter the market. We will assume that the rate at which job offers present themselves to workers depends on M_f so that we have $\lambda(M_f)$. Invert this function to write $M_f(\lambda)$. In equilibrium, the number of firms is determined by the free entry condition so that

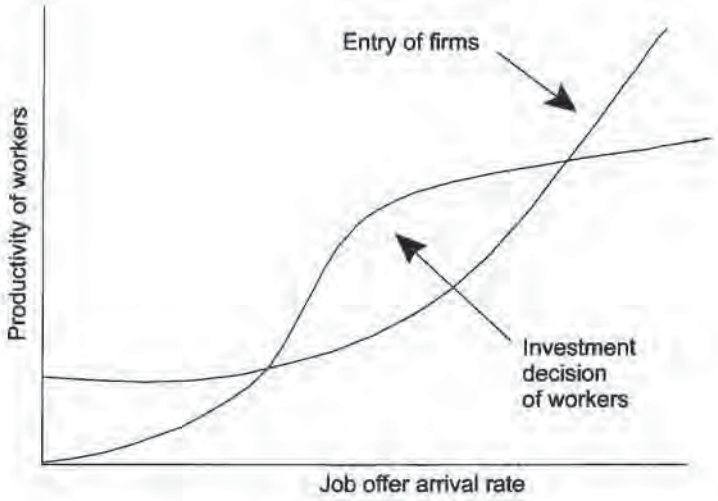


Figure 3.1 Multiple equilibria in the model of a ghetto.

$$\frac{\delta\lambda(p-b)}{M_f(\lambda)[\delta+\lambda]^2} = C_f \tag{3.13}$$

This gives λ as an increasing function of p with $\lambda \to \infty$ as $p \to \infty$. A possible outcome is given by the line marked "entry of firms" in figure 3.1. Inspection of figure 3.1 shows the possibility of multiple equilibria. As drawn, one can argue that only the high-level equilibrium is stable but one could always draw pictures with more than two stable equilibria. If this is the case then high-level equilibria will have high levels of human capital investment, high wages, high employment rates, and many firms. The differences in labor market outcomes can all be "explained" by differences in productivity but these differences in productivity are themselves the result of the poor expected labor market outcomes.

In this model, firms do not invest in the ghetto because of the poor "quality" of workers who live there. The residents do not invest in improving their skills because there are no jobs. A vicious circle is at work.

3.7 Conclusions

This chapter has considered the issue of efficiency in a number of theoretical models of oligopsonistic labor markets. It has one positive message: the free market must not be presumed efficient. But, beyond this, theory has been shown to provide little guidance about the direction of the

efficiency or policies that might be expected to improve matters. Employment may be too "high," or too "low." A minimum wage may raise employment or reduce it. If a theoretical paper claims a strong conclusion about the direction of inefficiency in the free market equilibrium, then this is almost certainly because they have not considered a rich enough model in the sense that there are not enough "marginal" decisions to be influenced by incentives. This chapter has introduced three such margins: elasticity in the supply of firms to the market, elasticity in the supply of workers, and endogenous recruitment intensity, and has shown how the nature of the inefficiency induced by each of them is rather different. As discussed in the introduction, there are other margins and these may be as important. We have introduced only enough to make the point that theory alone is going to be an unreliable guide to policy-making.

The rest of the book concentrates on positive rather than normative implications of employer market power. Even though the word "monopsony" conjures up emotive images of workers being exploited (in the sense of the word used by Hicks and Pigou) by employers, one should resist such temptations. Monopsony should simply be taken to mean that the supply of labor to the firm is not perfectly elastic. Although a recurrent theme is that the perspective of monopsony encourages one to be more open-minded about the likely impact of a range of policies than a strict believer in perfect competition would be inclined to be, the book is primarily concerned with what labor market phenomena can be better understood from the perspective of monopsony.

Appendix 3

Proof of Proposition 3.1

The derivative of the social surplus with respect to M is

$$\Omega'(M) = \frac{\delta\lambda'(M)(p-b)}{[\delta+\lambda(M)]^2} - C_f \tag{3.14}$$

At the free market level of M we have, using (3.2), that

$$\Omega'(M) = C_f\left[\frac{M\lambda'(M)}{\lambda(M)} - 1\right] < 0 \tag{3.15}$$

where the final inequality follows from the fact that $\lambda(M)$ is a strictly concave function of M. (3.15) implies that there are too many firms in the free market equilibrium and that the social surplus would be maximized by having fewer firms.

A binding minimum wage of w_m becomes the lowest wage offered in equilibrium. Hence, the profits made by the lowest-wage firm (and hence the equilibrium level of profits) is given by the left-hand side of

$$\frac{\delta\lambda(M)(p - w_m)}{M[\delta + \lambda(M)]^2} = C_f \tag{3.16}$$

and the equilibrium number of firms will solve (3.16). If this is to be equal to the socially efficient level of M (the level that solves $\Omega'(M) = 0$), it must be the case that

$$\frac{p - w_m}{p - b} = \frac{M\lambda'(M)}{\lambda(M)} \tag{3.17}$$

where the right-hand side should be evaluated at the efficient level of M: this is (3.4).

Proof of Proposition 3.2

Denote by $M_w(1 - u)G(w)$ the number of workers in equilibrium who are employed at a wage w or less:[6] this is the notation used previously in chapter 2.

Consider the flow of recruits to a firm that spends z on recruitment and offers a wage w. The fraction of total matches of workers to firms is given by the share of this firm in total recruitment activity, that is, by (z/Z). Of these matches only those involving employed workers in a job currently paying less than w or who are unemployed result in recruitment so that the flow of recruits to the firm is given by

$$R(w, z) = \frac{z}{Z}\left[\frac{\lambda(Z, M)}{M}(1 - u)G(w) + \frac{\lambda(Z, M)}{M}u\right] \tag{3.18}$$

Steady-state employment will be given by $N(w) = R/s(w)$ where $s(w)$ is the separation rate so that profits can be written as

$$\pi(w, z) = (p - w)\frac{(z/Z)[\lambda(Z, M)(1 - u)G(w) + \lambda(Z, M)u]}{M[\delta + \lambda(1 - F(w)]} - c(z)$$

$$\equiv \frac{z}{Z}\pi(w) - c(z) \tag{3.19}$$

where $\pi(w)$ is the term in (3.19) that does not involve z. (w, z) will be chosen by firms to maximize (3.19). For this profit maximization to be well defined, we obviously require that $c(\cdot)$ be convex, otherwise profits can be increased without bound. All combinations of (w, z) that are offered in equilibrium must yield the same level of profit. Maximizing

[6] These functions will also depend on $\{z(w), F(w)\}$ but this is suppressed to keep the notation simple.

(3.19) with respect to z while holding w constant leads to the conclusion that the optimal z will be a positive function of π and, by the envelope theorem, total profits must then be an increasing function of π. As all firms must make the same level of profits in equilibrium, $\pi(w)$ and, hence z, must be constant across firms. Given this fact, a comparison of (3.19) and (2.21) shows that, conditional on the chosen z, the distribution of wages must be the same as in the basic model derived in section 2.4.

Is the level of recruitment activity efficient? The social surplus can be written as

$$\Omega(z, M) = M_w[(1-u)p + ub] - M_f c(z) - M_f C_f$$

$$= M_w\left[p - \frac{\delta(p-b)}{[\delta + \lambda(z, M)]} - MC_f - Mc(z)\right] \tag{3.20}$$

The first-order condition for the derivative of Ω with respect to z is

$$\Omega_z(z, M) = \frac{\delta\lambda_z(z, M)(p-b)}{[\delta + \lambda(z, M)]^2} - Mc'(z) \tag{3.21}$$

where a subscript denotes a derivative with respect to that variable. At the free market level of z we have, using (3.6), that

$$\Omega_z(z, M) = Mc'(z)\left(\frac{z\lambda_z(z, M)}{\lambda(z, M)} - 1\right) < 0 \tag{3.22}$$

where the final inequality follows from the fact that $\lambda(z, M)$ is assumed to be a concave function of z. (3.22) implies that there is excessive recruitment activity in the free market equilibrium and that the social surplus would be increased by reducing recruitment activity.

A binding minimum wage of w_m becomes the lowest wage offered in equilibrium. Hence, the level of z chosen by firms in the free-market equilibrium will be given by

$$\frac{\delta\lambda(z, M)(p - w_m)}{M[\delta + \lambda(z, M)]^2} = zc'(z) \tag{3.23}$$

If this is to be equal to the socially efficient level of z, from (3.21) it must be the case that

$$\frac{p - w_m}{p - b} = \frac{z\lambda_z(z, M)}{\lambda(z, M)} \tag{3.24}$$

where the right-hand side should be evaluated at the efficient level of z.

Proof of Proposition 3.3

Differentiate (3.10) to yield

$$\frac{\partial V(w)}{\partial w} = \frac{1}{\delta + \lambda(1 - F(w))} \tag{3.25}$$

Now integrating the integral term in (3.8) by parts and using (3.25), we can obtain

$$\int_{w_m} [V(x) - V^u] dF(x) = [V(w_m) - V^u] + \int_{w_m} \frac{\partial V(x)}{\partial x} [1 - F(x)] dx \tag{3.26}$$

Now, using (3.10), we have that

$$[V(w_m) - V^u] = \frac{w_m - b}{\delta + \lambda} \tag{3.27}$$

Using (3.25), (3.26) and (3.27) in (3.8) then leads to the following expression for V^u:

$$\delta_r V^u = b + \frac{\lambda}{\delta + \lambda}(w_m - b) + \int_{w_m} \frac{\lambda[1 - F(x)]dx}{\delta + \lambda(1 - F(x))} \tag{3.28}$$

Integrating the final integral term in (3.28) by parts yields

$$\delta_r V^u = b + \frac{\lambda}{\delta + \lambda}(w_m - b) + \int_{w_m} (x - w_m) \frac{\delta\lambda f(x)dx}{[\delta + \lambda(1 - F(x))]^2} \tag{3.29}$$

Using (2.20), this can be written as

$$\delta_r V^u = b + \frac{\lambda}{\delta + \lambda}(w_m - b) + M \int_{w_m} (x - w_m) N(x) f(x) dx \tag{3.30}$$

Using the expression for the aggregate non-employment rate in (2.17), this can be written as

$$\delta_r V^u = b + \frac{\lambda}{\delta + \lambda}(w_m - b) + \frac{\lambda}{\delta + \lambda}[E(w) - w_m]$$

$$= b + \frac{\lambda}{\delta + \lambda}(E(w) - b) \tag{3.31}$$

Now, in equilibrium, it must be the case that

$$(p - w)N(w) = (p - w_m)N(w_m) \tag{3.32}$$

for all offered wages. Taking expectations, using (2.17) and (2.20) for the lowest-wage firm and rearranging, yields

$$E(w) = p - \frac{\delta}{\delta + \lambda}(p - w_m) \tag{3.33}$$

Substituting this into (3.31) leads to

$$\delta_r V^u = b + \frac{\lambda}{\delta + \lambda}(w_m - b) + \frac{\lambda^2}{(\delta + \lambda)^2}(p - w_m) \tag{3.34}$$

In the free market equilibrium we must have $V^u = V^n$ and $w_m = b$. Using (3.9) and (3.34), the free market entry condition can be written as

$$\frac{\lambda(M)^2}{(\delta + \lambda(M))^2}(p - b) = C_w \tag{3.35}$$

To derive the social surplus, assume that wages are always equal to p so all the social surplus goes to workers. Then, the value functions, (3.8) and (3.10), become

$$\delta_r V^u = b + \lambda[V(p) - V^u] \tag{3.36}$$

$$\delta_r V(p) = p + \delta_u[V^u - V(p)] \tag{3.37}$$

which implies that

$$\delta_r V^u = b + \frac{\lambda(p - b)}{\delta + \lambda} \tag{3.38}$$

Now the total social surplus is $(V^u - V^n)M_w$ as V^u is the lifetime expected utility of a worker entering the labor market. This, using the fact that $M = M_f/M_w$, yields

$$\Omega(M) = \left[\frac{\lambda(M)}{\delta + \lambda(M)}(p - b) - C_w\right]\frac{M_f}{M} \tag{3.39}$$

Differentiating (3.42), we obtain

$$\Omega'(M) = -\left[\frac{\lambda(M)}{\delta + \lambda(M)}(p - b) - C_w\right]\frac{M_f}{M^2} + \frac{\delta\lambda_M(M)}{[\delta + \lambda(M)]^2}(p - b)\frac{M_f}{M} \tag{3.40}$$

which, using (3.35) to eliminate $(p - b)$ in (3.40) yields, after some rearrangement,

$$\Omega'(M) = \frac{\delta M_f C_w}{\lambda M^2}[\varepsilon_{\lambda M} - 1] < 0 \tag{3.41}$$

so that there are too few workers entering the labor market. This proves part 1 of Proposition 3.3.

Now consider part 2. If we set $\Omega'(M) = 0$, then (3.40) can be written as

$$\frac{\lambda(M)}{\delta + \lambda(M)}(p - b) - C_w = \frac{\delta M\lambda_M(M)}{[\delta + \lambda(M)]^2}(p - b) \tag{3.42}$$

If $V^u = V^n$, then from (3.34) and substituting the expression for C_w, the equilibrium condition for efficiency can be written as

$$\frac{\lambda(M)}{\delta+\lambda(M)}(p-w_m)-\frac{\lambda(M)^2}{[\delta+\lambda(M)]^2}(p-w_m)=\frac{\delta M\lambda_M(M)}{[\delta+\lambda(M)]^2}(p-b)=C_w \tag{3.43}$$

which, on rearrangement, leads to (3.11).

Proof of Proposition 3.4

In proving this proposition, it is helpful to first prove the following lemma on the equilibrium of the model.

Lemma 3.1

1. *If a firm pays a wage w, the supply of labor to it, $N(w)$ will be given by*

$$N(w;F)=\frac{\delta\lambda H(w)}{M[\delta+\lambda(1-F(w))]^2} \tag{3.44}$$

2. *The lowest wage offered in the free market equilibrium, w_0, is the solution to*

$$w_0=\text{argmax}(p-w)H(w) \tag{3.45}$$

3. *The equilibrium level of profits, π^*, is given by*

$$\pi^*=\frac{\delta\lambda(p-w_0)H(w_0)}{M[\delta+\lambda]^2} \tag{3.46}$$

4. *The equilibrium wage offer distribution is found by solving*[7]

$$\frac{(p-w)H(w)}{[\delta+\lambda(1-F(w))]^2}=\frac{(p-w_0)H(w_0)}{[\delta+\lambda]^2} \tag{3.47}$$

Proof. For workers with a reservation wage b, the non-employment rate will be

$$u(b)=\frac{\delta}{\delta+\lambda[1-F(b)]} \tag{3.48}$$

as only a fraction $[1-F(b)]$ of job offers are acceptable.

[7] One might be concerned that the solution $F(w)$ to (3.47) need not be a legitimate distribution function as it is possible that it decreases for some w. If this occurs, the equilibrium wage distribution is the upper envelope of the solution to (3.47). But log-concavity of $H(w)$ is sufficient to ensure the solution to (3.47) is a legitimate distribution function.

Define $M(w)$ to be the number of workers with a wage less than w. By analogy to the argument used in deriving Proposition 2.3, we must have

$$[\delta + \lambda(1 - F(w))]M(w) = \lambda M_w \int^w [F(w) - F(b)]u(b)h(b)db \quad (3.49)$$

as unemployed workers with reservation wage b accept jobs paying w or less at a rate $\lambda[F(w) - F(b)]$. The supply of labor to an individual firm can then be written as

$$[\delta + \lambda(1 - F(w))]N(w; F) = \frac{\lambda}{M_f}\left[M(w) + M_w \int^w u(b)h(b)db\right] \quad (3.50)$$

Substituting (3.48) and (3.49) into (3.50) leads, after some rearrangement to (3.44) which proves part 1.

Using (3.44), profits can be written as

$$\pi(w; F) = \frac{\delta\lambda(p - w)H(w)}{M[\delta + \lambda(1 - F(w))]^2} \quad (3.51)$$

Suppose the lowest wage is above w_0 as defined by (3.45). Then profits in the lowest-wage firm can be increased by cutting wages to w_0 as $F(w)$ will still be equal to zero. Similarly if the lowest wage is below w_0 then profits can be increased by increasing the wage paid as $(p - w)H(w)$ must rise and $F(w)$ cannot fall.

In equilibrium, all offered wages must yield the same level of profit. Using (3.51) and (3.46) leads to (3.47). This completes the proof of the lemma.

The proof of Proposition 3.4 is now very simple. As long as some offered wage is below the highest value of b, some efficient matches will not be consummated. A minimum wage of p guarantees that this does not happen.

Proof of Proposition 3.5

Suppose the social planner is choosing the number of firms. They obviously want to maximize employment for all those with $p > b$ so that the social surplus can be written as

$$\Omega(M) = \left[\frac{\lambda(M)}{\delta + \lambda(M)} \int (p - b)h(b)db - MC_f\right]M_w \quad (3.52)$$

and the first-order condition for the efficient level of M can be written as

$$\Omega'(M) = \left[\frac{\delta\lambda'(M)}{[\delta + \lambda(M)]^2} \int (p - b)h(b)db - C_f\right]M_w = 0 \quad (3.53)$$

Using (3.46), we have that at the free market equilibrium, C_f is equal to the level of profits given by (3.47). Hence, the derivative of the social surplus with respect to M is

$$\Omega'(M) = \frac{\delta\lambda(M)}{M[\delta + \lambda(M)]^2}\left[\varepsilon_{\lambda M}\int(p - b)h(b)db - (p - w_0)H(w_0)\right]M_w \tag{3.54}$$

The sign of this is ambiguous. One the one hand, $\varepsilon_{\lambda M} < 1$ which tends to make (3.54) negative implying there are too many firms. On the other hand,

$$(p - w_0)H(w_0) = \int^{w_0}(p - w_0)h(b)db < \int^{w_0}(p - b)h(b)db$$

$$< \int(p - b)h(b)db \tag{3.55}$$

which tends to make (3.54) positive. This proves the first part of the proposition.

If the unemployment rate of workers with reservation wage b is $u(b)$, then total surplus per worker can be written as

$$\Omega = \int(p - b)h(b)db - \int(p - b)u(b)h(b)db - MC_f \tag{3.56}$$

Now, if the lowest wage paid is w_m (that might be determined by a binding minimum wage), then using (3.48), this can be written as

$$\Omega = [p - E(b)] - \frac{\delta}{\delta + \lambda}\int^{w_m}(p - b)h(b)db - \int_{w_m}^{\hat{w}}\frac{\delta(p - b)h(b)db}{\delta + \lambda[1 - F(b)]}$$

$$- \int_{\hat{w}}(p - b)h(b)db - MC_f \tag{3.57}$$

where $\hat{w}$ is the highest wage.

Now, from (3.47) we can eliminate $F(b)$ from the third term on the right-hand side of (3.57) and write Ω as

$$\Omega = [p - E(b)] - \frac{\delta}{\delta + \lambda}\int^{w_m}(p - b)h(b)db$$

$$- \frac{\delta\sqrt{(p - w_m)H(w_m)}}{\delta + \lambda}\int_{w_m}^{\hat{w}}\frac{(p - b)h(b)db}{\sqrt{(p - b)H(b)}}$$

$$- \int_{\hat{w}}(p - b)h(b)db - MC_f \tag{3.58}$$

and the free entry condition can be written as

$$\frac{\delta\lambda(M)(p-w_m)H(w_m)}{M[\delta+\lambda(M)]^2}=C_f \tag{3.59}$$

From the free entry condition, we have that the effect of the minimum wage on the number of firms is given by

$$\left[\frac{\lambda'(M)}{\lambda(M)}-\frac{1}{M}-\frac{2\lambda'(M)}{\delta+\lambda}\right]\frac{\partial M}{\partial w_m}=-\frac{\Psi'(w_m)}{\Psi(w_m)} \tag{3.60}$$

where $\Psi(w)\equiv(p-w)H(w)$. Note that a minimum wage set at w_0 (i.e., just binds) will have no effect on the number of firms as $\Psi'(w_0)=0$. But a minimum wage above w_0 will have $\Psi'(w_m)<0$.

Now, differentiating (3.58), we have that

$$\begin{aligned}\frac{\partial\Omega}{\partial w_m}=&\left[\frac{\delta}{[\delta+\lambda]^2}\int_{w_m}^{\hat{w}}(p-b)h(b)db\right.\\&\left.+\frac{\delta\sqrt{(p-w_m)H(w_m)}}{[\delta+\lambda]^2}\int_{w_m}^{\hat{w}}\frac{(p-b)h(b)db}{\sqrt{(p-b)H(b)}}\right]\lambda'(M)\frac{\partial M}{\partial w_m}\\&-\frac{\delta\Psi'(w_m)}{2\sqrt{\Psi(w_m)}[\delta+\lambda]}\int_{w_m}^{\hat{w}}\frac{(p-b)h(b)db}{\sqrt{(p-b)H(b)}}-C_f\frac{\partial M}{\partial w_m}\end{aligned} \tag{3.61}$$

Using (3.60) to eliminate the term in $\Psi'(w_m)$, one can write this as

$$\begin{aligned}\frac{\partial\Omega}{\partial w_m}=\frac{\partial M}{\partial w_m}&\left[\frac{\delta\lambda'(M)}{[\delta+\lambda]^2}\int_{w_m}^{\hat{w}}(p-b)h(b)db\right.\\&\left.+\frac{\delta\sqrt{\Psi(w_m)}[\varepsilon_{\lambda M}-1]}{2M[\delta+\lambda]}\int_{w_m}^{\hat{w}}\frac{(p-b)h(b)db}{\sqrt{(p-b)H(b)}}-C_f\right]\end{aligned} \tag{3.62}$$

If the minimum wage just binds, all these terms are zero as they all involve $\partial M/\partial w_m$ which is zero at that point. But, for strictly binding minimum wages we have that $(\partial M/\partial w_m)<0$ so that the sign of (3.62) is determined by the sign of the terms in the square brackets on the right-hand side. The first term is positive and the others negative, making the overall sign ambiguous.

Proof of Proposition 3.6

From Proposition 3.4 and (3.34) we know that, for the case where the lowest wage is equal to b, the value of being non-employed can be written as

$$\delta_r V^u = b + \left(\frac{\lambda}{\delta + \lambda}\right)^2 (p - b) \tag{3.63}$$

Maximizing $\delta_r V^u - c(p)$ then leads to (3.12).

4

The Elasticity of the Labor Supply Curve to an Individual Firm

THE single most important idea in this book is that the wage elasticity of the labor supply curve (ε_{Nw} in the notation of previous chapters) is not infinite or close to it. Hence, the most direct way to establish the existence of employer market power over its workers is to estimate the wage elasticity of the labor supply curve facing the firm. Studies of this elasticity are few and far between: one might cite Reynolds (1946a), Nelson (1973), Sullivan (1989), Machin et al. (1993), Boal (1995), Beck et al. (1998), Staiger et al. (1999), and Falch (2001) as an almost complete list. This lack of literature contrasts with entire books written about the demand for labor or the supply of labor by individuals and with the literature on industrial organization on estimating the extent of product market power (for a survey, see Bresnahan 1989). It is testament to the faith that most labor economists have that $\varepsilon_{Nw} = \infty$. But, given the paucity of the literature, this is nothing but faith and some of us might want some evidence that ε_{Nw} is "high" if not infinite.

The plan of this chapter is as follows. The first section discusses the problems of using correlations between wages and employment to estimate the wage elasticity of the labor supply curve facing the firm. We review the literature on the employer size–wage effect, arguing (in the third section) that the evidence suggests that part (though not all) of the employer–size wage effect is the result of an upward-sloping labor supply curve to the individual firm. We argue that, in the absence of good instruments in the form of firm-level demand shocks, it is hard to get a good estimate in this way of the wage elasticity of the labor supply curve facing an individual employer. The fourth section then discusses an alternative method based on using a dynamic monopsony model and estimating the elasticity of separations and recruits with respect to the wage. Finally, the chapter discusses estimates derived from more structural estimation of equilibrium search models.

The main conclusion is that the elasticity of the labor supply curve facing the firm does not seem to be close to infinite but that it is hard to get a very precise estimate of it. An estimate in the region of 2–5 seems to be reasonable. These estimates imply that employers have sizeable amounts of monopsony power: even an elasticity of 5 implies

that wages will (using (2.3)) be 17% below the marginal revenue product.

4.1 The Employer Size–Wage Effect

A simple-minded approach to estimating the elasticity of the labor supply curve facing the firm would be to simply regress the log of the wage that the firm pays on the log of employment plus any other variables that might be thought to be important controls. If the market is perfectly competitive we should find a coefficient of zero on employment whereas monopsony would predict it to be positive and the size of the coefficient would give an estimate of the extent of the monopsony power possessed by the firm (as it estimates $\varepsilon = (1/\varepsilon_{Nw})$). There is a large empirical literature that estimates this type of regression and finds a significant positive relationship between wages and employment—what is commonly known as the employer size–wage effect (ESWE). However, an upward-sloping labor supply curve is not the only explanation proposed for the ESWE (for surveys, see Brown and Medoff 1989; Brown et al. 1990; or Oi and Idson 1999) and plausible alternatives need to be considered. Indeed, much of the literature on the ESWE does not even consider an upward-sloping labor supply curve to an individual employer as a possible explanation.[1] This is in spite of the fact that an innocent might think that the first hypothesis an economist would investigate when observing a positive relationship between a price (the wage) and a quantity (employment) is that it is a supply curve. However, Brown and Medoff (1989: 1056) do not manage to identify the cause of the employer size–wage effect and conclude that "our analysis leaves us uncomfortably unable to explain it." Here, we argue that monopsony can fill that void.

Consider a very simple stripped-down model. Assume that firm i has a revenue function which is given by

$$Y_i = \frac{1}{1-\eta} A_i N_i^{1-\eta} \tag{4.1}$$

where A_i is a shock to the marginal revenue product of labor (MPRL) curve. On the supply side of the labor market, assume that the wage that the firm pays is given by

$$w_i = B_i N_i^{\varepsilon} \tag{4.2}$$

[1] For example, the otherwise very thorough and even-handed survey of the relationship between employer size and wages of Brown and Medoff (1989) does not mention monopsony at all with the possible exception that there is some discussion of the rather involved and convoluted "labor pools" model of Weiss and Landau (1984) which could be seen as a sort of monopsony model.

where B_i is a shock to the supply curve. These supply shocks could represent differences in local labor market conditions (because of skill or regional differences) or differences in the attractiveness of non-wage attributes in different firms. We are interested in obtaining a consistent estimate of ε, the inverse elasticity of the labor supply curve facing the firm.

The firm will choose a level of employment where the MPRL equals the marginal cost of labor so that the chosen employment level will satisfy

$$A_i N_i^{-\eta} = (1 + \varepsilon) B_i N_i^{\varepsilon} \tag{4.3}$$

or, in log-linear form

$$\log(N_i) = \frac{1}{\varepsilon + \eta}[a_i - b_i - \ln(1 + \varepsilon)] \tag{4.4}$$

where $a = log(A)$ and $b = log(B)$. The chosen wage will be given by

$$\log(w_i) = \frac{1}{\varepsilon + \eta}[\varepsilon a_i + \eta b_i - \varepsilon \ln(1 + \varepsilon)] \tag{4.5}$$

(4.4) and (4.5) are easy to understand. Positive shocks to the MRPL cause employment and wages to rise, although there is only an effect on wages to the extent that the employers do have some labor market power ($\varepsilon > 0$). Positive shocks to the labor supply curve cause employment to fall and wages to rise.[2]

Now make the following assumptions about the observability of the shocks (a,b):

$$a_i = \beta_a x_i + \nu_{ai}$$

$$b_i = \beta_b x_i + \nu_{bi} \tag{4.6}$$

where x is a set of explanatory variables observable to the researcher. In the interests of notational simplicity assume that the same variables affect both a and b. Of course, a particular variable can be constrained to affect only demand or supply shocks by imposing the restriction that its coefficient in the other equation is zero. Assume that the shocks ν are independent of x and jointly normally distributed with mean zero and covariance matrix Σ. Denote by σ_a^2 the variance of ν_a, σ_b^2 the variance of ν_b and σ_{ab} the covariance between ν_a and ν_b.

Now consider how one might set about estimating ε. First, one might think about estimating by OLS the relationship between the log wage and

[2] The model used here is simple in the sense that it uses a static labor supply curve and assumes the employer can only use the wage to influence the supply of labor to the firm: Appendix 4B presents analyses of dynamic labor supply curves and what happens if the generalized model of monopsony of section 2.3 is used.

log employment controlling for other factors thought to be relevant (the x variables in our notation). If one thought of the aim of this as being to estimate a supply curve facing the firm, then one might think of including only those x variables which affect supply (i.e., exclude those affecting only demand). However, many researchers have not thought of their purpose as estimating the supply curve facing a firm so have not used this argument for excluding some variables. For example, when estimating earnings functions, it is common practice to include employment as simply another regressor and the researcher does not think of their aim as being to estimate a supply curve to the individual firm. So, again, we would like to have some idea of the consequences of using this "kitchen sink" approach or using the regressions of others to estimate ε.

A regression of $\log(w_i)$ on $(\log(N_i), x_i)$ estimates $E(\log(w_i) \mid \log(N_i), x_i)$. The following proposition tells us what we would expect to find.

Proposition 4.1. *Running a regression of* $\log(w_i)$ *on* $(\log(N_i), x_i)$ *estimates*

$$E(\log(w_i) \mid \log(N_i), x_i) = (\varepsilon + \rho(\varepsilon + \eta))\log(N_i) + \rho \ln(1 + \varepsilon) + (\beta_b - \rho(\beta_a - \beta_b))x_i \quad (4.7)$$

where

$$\rho \equiv \frac{\sigma_{ab} - \sigma_b^2}{\sigma_a^2 + \sigma_b^2 - 2\sigma_{ab}} \quad (4.8)$$

Proof. See Appendix 4A.

(4.7) says that the kitchen sink approach will only give an unbiased estimate of ε if $\rho = 0$ which implies that ν_a can be written as ν_b plus some uncorrelated noise. A special case of this is when there are no unobserved supply shocks. In this case, all firms have the same labor supply curve (conditional on x) and variation in N caused by unobserved demand shocks will trace out the labor supply curve. In any other situation one will end up with a biased estimate of ε. If ν_a and ν_b are uncorrelated, the estimate of ε has a downward bias as (4.8) then implies that $\rho < 0$. Intuitively unobserved shifts in the labor supply curve cause wages and employment to move in opposite directions making the slope of the supply curve seem less positive than it really is.

(4.7) can also be used to understand the arguments that the estimated employer size–wage effect overstates the true value of ε (which must be the case if one believes that labor markets are competitive and $\varepsilon = 0$). The two main arguments are unobserved labor quality and compensating wage differentials. One would expect high-quality workers to have a

high level of a (as their productivity is high) and a high level of b as b will partly reflect the wages paid by other firms. So, one would expect unobserved labor quality to result in $\sigma_{ab} > 0$. But, from (4.7) and (4.8) this is not sufficient to imply an upward bias: for that we require $\sigma_{ab} > \sigma_b^2$ (or, equivalently that the expectation of $(a - b)$ is increasing in b). If, for example proportional differences in a are reflected in proportional differences in b, then this is exactly the situation in which we obtain an unbiased estimate of ε. However, the fact that workers with high levels of observable skills are more likely to work in large firms does suggest that the condition $\sigma_{ab} > \sigma_b^2$ might be satisfied if the same is true of unobserved skills.[3] One can also understand compensating differentials as a positive correlation between a and b (as the disamenity must have some positive effect on productivity) so that the previous discussion is also relevant for this case. Let us consider the evidence that all of the ESWE can be explained through these effects.

4.2 Competing Explanations for the Employer Size–Wage Effect

Among the strongest contenders for an explanation of the size–wage effect (apart from an upward-sloping labor supply curve) are

- unobserved worker quality;
- compensating wage differentials;
- rent sharing.

All of these possibilities are discussed by Brown and Medoff (1989) in their survey and much of the discussion here is similar.

Table 4.1 presents some basic information on the size–wage effect for the United States (from the April 1993 Contingent Worker Survey (CWS) supplement to the CPS) and the United Kingdom (from the LFS). In both countries, information on employer size is banded, the bands used differing slightly in the two countries. The measure of employer size used is workplace size although the CPS also has information on firm size (which also seems to have a positive impact on wages independent of workplace size). The distribution of workers by establishment size reported in the column headed "sample percentages" is remarkably similar in both countries, the median worker being in a workplace with slightly more than 50 workers.

The column marked (1) for both countries presents estimates of the size–wage effect when there are no other controls in a wage equation. The

[3] In the United States, 12.6% of college graduates work in plants with less than 10 employees and 34% in plants with more than 250. For those who are not college graduates, the figures are 26% in plants with less than 10 workers and 24% in those with more than 250. The United Kingdom is similar.

TABLE 4.1
The Employer Size–Wage Effect in the United States and the United Kingdom

Number of Employees	*United States (April 1993 CPS)*			*United Kingdom (LFS)*		
	Sample Percentage	*(1)*	*(2)*	*Sample Percentage*	*(1)*	*(2)*
1–10	19.9	−0.226 (0.022)	−0.118 (0.019)	18.4	−0.268 (0.004)	−0.151 (0.004)
11–19	13.5	−0.044 (0.024)	−0.015 (0.020)	9.5	−0.097 (0.005)	−0.040 (0.004)
20–24				4.4	−0.048 (0.007)	−0.018 (0.005)
25–49	14.9	0	0	12.5	0	0
50–99	13	0.098 (0.024)	0.067 (0.020)			
100–249	13.1	0.163 (0.024)	0.098 (0.020)	55.2	0.157 (0.006)	0.073 (0.006)
250+	27.6	0.289 (0.020)	0.182 (0.018)			
Other controls	n.a.	No	Yes	n.a.	No	Yes
Number of observations	7854	7854	7854	220868	220868	220868
R^2	n.a.	0.1	0.36	n.a.	0.07	0.42

Notes.
1. The dependent variable is the log of the hourly wage. The other controls included are marital status, children, experience, tenure (all interacted with gender), region, race, and (for the LFS) month dummies. Standard errors are reported in parentheses.

reference category is a workplace with 25–49 workers. The gap in average log wages between the smallest (1–10 employees in both the CPS and the LFS) and the largest workplaces (250+ employees in the United States, 50+ in the United Kingdom) is 0.515 log points in the United States and 0.425 in the United Kingdom. Introducing controls (the columns marked (2)), reduces the magnitude of the effect to approximately half. In both countries it is the introduction of controls for education, experience, and tenure that has the biggest effect in reducing the size–wage effect.

The reduction in the size–wage effect when controls are introduced suggests that part of the raw size–wage differential can be explained by differences in worker quality. But, are estimates that control for worker quality inevitably better than those that do not? To answer this question, let us return to (4.7).

Controlling for labor quality is likely to reduce the unobserved parts of the labor supply and MRPL equations (ν_a and ν_b in (4.6)) so that σ_a^2 and σ_b^2 are reduced in magnitude as, presumably, is σ_{ab}. For a single equation, the standard formula for the extent of omitted variable bias might lead one to believe that the bias is reduced and the resulting estimates are "better." But, matters are more subtle in a simultaneous equations model as the unexplained part of the regressor (here, employer size) is also reduced by the introduction of controls. In fact, (4.7) and (4.8) show that it is (σ_a^2/σ_b^2) and the correlation coefficient between ν_a and ν_b that determines the bias so it is relative, not absolute, variances that are important in determining whether the introduction of controls reduces or increases the bias.

Why does the coefficient on employer size tend to fall when other controls are included? If one thinks that it is mostly "labor quality" and regional variables that induce the correlation between ν_a and ν_b, then we might expect that the correlation between these residuals falls when we improve our controls for these variables. If (for want of a better reason) we assume the relative variances are constant then, using (4.8), it is simple to check that a fall in the correlation between ν_a and ν_b reduces the coefficient on employer size. But, is this reduced coefficient a better estimate of the true value of ε? Not necessarily: as the previous discussion has made clear, reducing the correlation between ν_a and ν_b to zero will lead to an underestimate of the true labor supply elasticity (set $\sigma_{ab} = 0$ in (4.8)). Hence, one should not leap to the conclusion that controlling for labor quality inevitably leads to a better estimate of ε.

However, in spite of this, the reduction in the size–wage effect when controls are introduced does suggest that part of the raw size–wage differential might be explained by differences in worker quality. As a large part of worker quality is unobserved, this has led some to argue that all of the

size–wage effect might be explained by differences in worker quality. A common way of controlling for unobserved worker quality is to use panel data to estimate a fixed-effects model: that is, regress changes in wages on changes in employer size. Neither the CPS nor the LFS used above are panel data sets so we use the BHPS for the United Kingdom to investigate this. The results are reported in table 4.2. To make the presentation of the results simple, estimates of a simple elasticity of wages with respect to employer size are presented. Log employer size is computed using the mid-points of the reported bands: more sophisticated attempts to predict employer size conditional on characteristics made little difference to the results.

The first four rows present estimates of the elasticity from the CPS and LFS both with and without controls. Without controls the elasticity of wages with respect to employer size is 0.11 in the United States and 0.14

TABLE 4.2
The Elasticity of Wages with Respect to Employer Size

	Country (Data)	*Sample*	*Other Controls*	*Elasticity (SE)*	*Number of Observations*	R^2
(1)	US (CPS)	Cross-section	No	0.108 (0.004)	7854	0.10
(2)	US (CPS)	Cross-section	Yes	0.064 (0.003)	7854	0.36
(3)	UK (LFS)	Cross-section	No	0.145 (0.002)	220868	0.07
(4)	UK (LFS)	Cross-section	Yes	0.074 (0.001)	220868	0.42
(5)	UK (BHPS)	Cross-section	No	0.086 (0.002)	13365	0.09
(6)	UK (BHPS)	Cross-section	Yes	0.047 (0.002)	13365	0.49
(7)	UK (BHPS)	Panel	Yes	0.013 (0.002)	13813	0.02
(8)	UK (BHPS)	Panel movers	Yes	0.035 (0.007)	1340	0.11
(9)	UK (LFS)	Dual job holders	No	0.037 (0.007)	5342	0.01

Notes.
1. The dependent variable is the log of the hourly wage. Employer size is coded as the mid-points of the relevant bands with the open-ended top category being coded as twice the lower bound. The elasticity is the coefficient on the log of the employer size variable.
2. The other controls included are marital status, children, experience, tenure (all interacted with gender), region, race, and (for the LFS) month dummies. Standard errors are reported in parentheses.

in the United Kingdom. Introducing controls reduces the estimates by 40–50% to 0.064 in the United States and 0.074 in the United Kingdom. The fifth and sixth rows estimate the elasticity using the BHPS: the elasticity here is smaller than that found in the LFS. The seventh row estimates an equation for wage growth including change in employer size as a regressor. The estimated elasticity drops to 0.013 although it remains significantly different from zero. This might suggest that controlling for unobserved worker quality makes most of the size–wage effect disappear. However, there is good reason to think that this is an understatement of the true elasticity as the employer size variable is likely to have a lot of measurement error (think of the difficulty in answering a question about the size of your workplace)[4] and this measurement error will be compounded when we estimate a model in first-differences. To get some idea of the extent of the problem, we compared worker responses to the employer size question in the BHPS to management responses to a similar question in the 1998 UK Workplace Employee Relations Survey (WERS).[5] In the BHPS only 63% of workers who did not change jobs reported their employer being in the same size class as one year ago: for WERS the (employee-weighted) figure is 88%. For the largest workplaces (those with 1000+ workers) 97% of the WERS sample reported being in the same category the previous year and the remaining 3% were in the next category (500–999 workers). In the BHPS, only 72% reported having 1000+ employees previously and 10% reported their employer previously having less than 200 employees. It is clear that there is a lot of measurement error in reported changes in employer size from employee data. This measurement error in the change in employer size is likely to be less important for workers who change jobs as the signal to noise ratio is likely to be higher. The eighth row of table 4.2 estimates a wage growth model on a sample of movers—the estimated elasticity rises to 0.037. Similar results are reported on US data sets by Brown and Medoff (1989).

[4] Measurement error is another reason why estimates that include controls may be worse than those that do not as was originally pointed out by Griliches (1977). Intuitively, the fraction of the variation in the employer size variable that is measurement error after introducing controls is likely to be higher. In the present context, measurement error should be interpreted broadly to mean any transitory shocks to employment. For example, it may be that wages are related to a long-run measure of employer size because there are pressures which limit the extent of variation in wages (see chapter 5 for a discussion of this) but that there are year-to-year variations in employment that do not get reflected in wages. This will have the same effect as measurement error on the estimated ESWE.

[5] WERS reports the actual level of employment now and a year ago, so we converted this to the size classes used in the BHPS. As WERS only reports weights which can be used to gross to the population of workplaces with 10+ employees, we also restricted this analysis to workers who report 10+ employees in the BHPS.

TABLE 4.3
Job Mobility and Employer Size

	(1)	(2)	(3)
Log employer size	−0.054 (0.008)	−0.038 (0.009)	−0.023 (0.009)
Log wage			−0.269 (0.042)
Other controls	No	Yes	Yes
Number of observations	13928	13886	13886
Pseudo-R^2	0.005	0.1	0.11

Notes.
1. The data set used is the BHPS. The sample is all of those in continuous employment between one interview and the next. The dependent variable takes the value 1 if the individual changed jobs and 0 if they did not: a probit model is estimated.
2. The other controls are sex, race, education, experience, tenure, region, marital status, children, and year dummies.

The LFS offers another way to control for individual fixed effects as, for those who have more than two jobs, it asks questions about earnings and employer size for both jobs.[6] The ninth row of table 4.2 regresses the difference in log wages on the difference in log workplace size; the estimated elasticity is 0.037, identical to that obtained from the BHPS movers. These estimates suggest that controlling for worker quality does reduce the size–wage effect but it remains significantly different from zero, implying a gap of about 10% in wages between the 75th and 25th percentile of workplace size: this is similar to the magnitudes reported by Brown and Medoff (1989).

One other hypothesis to explain the size–wage effect is that it is the result of a compensating wage differential. It may be that people dislike working in large firms per se or that other working conditions tend to be worse in large firms. Observed indicators of working conditions do not suggest that large firms tend to be worse places to work but these indicators are far from perfect. Perhaps the simplest way to test the compensating wage differentials hypothesis is to examine quit rates. If the size–wage effect is simply a compensating wage differential, utility would be equalized across firms of different sizes so there is no reason to believe that quit rates would differ by firm size. In fact, workers are much less likely to leave large employers. Table 4.3 presents some evidence from the BHPS on this point. The sample is those in continuous employment where the dependent variable takes the value 1 if the individual changed jobs and 0 if they did not. In column (1) we simply estimate a probit model with the log of employer size as a regressor. It

[6] Lemieux (1998) was the first to use this approach as a way to estimate the union wage mark-up.

is significantly negative. The second column then introduces some extra controls. The impact of employer size is reduced but still significant. Finally, the third column also introduces the log wage. It is not clear whether the wage should be included or not as the impact of employer size on quits should be through the wage. Again, the coefficient on employer size is reduced but remains significantly different from zero. This evidence suggests that workers are better off in large firms, which suggests that the size–wage correlation cannot be explained solely by compensating wage differentials.

The evidence discussed so far suggests that the size–wage effect cannot be readily explained by a competitive model of the labor market. However, this does not prove that an upward-sloping supply of labor to the individual employer is the correct explanation: there are other potential non-competitive explanations, for example, efficiency wage or rent sharing theories. While efficiency wage theories are often too vague to test, rent sharing is a tighter idea. The hypothesis is that workers manage to get a share of the rents and successful firms with large rents tend to have high levels of employment. If this is the case, we would expect that workers are better at extracting a share of the rents when they are unionized than when they are not so we would expect to see a larger employer size–wage effect in the union sector. In fact, the opposite is the case. Table 4.4 shows that in the CPS, the LFS and the BHPS the size–wage effect is 3–4 times larger in the non-union sector than the union sector (for similar findings, see Brown and Medoff 1989,

TABLE 4.4
The ESWE in Union and Non-union Sectors

	Country (Data)	*Sample*	*Other Controls*	*Elasticity (SE)*	*Number of Observations*	R^2
(1)	US (CPS)	Union cross-section	Yes	0.019 (0.008)	1231	0.30
(2)	US (CPS)	Non-union cross-section	Yes	0.067 (0.004)	6623	0.37
(3)	UK (LFS)	Union cross-section	Yes	0.017 (0.003)	22737	0.42
(4)	UK (LFS)	Non-union cross-section	Yes	0.086 (0.003)	24135	0.4
(5)	UK (BHPS)	Union cross-section	Yes	0.018 (0.002)	6619	0.47
(6)	UK (BHPS)	Non-union cross-section	Yes	0.077 (0.003)	6746	0.49

Notes.
1. As for table 4.2.

Mellow 1982).[7] This seems strong evidence against the rent sharing hypothesis; Green et al. (1996) provide a more extensive discussion of this for UK data.

This section has discussed competitive and rent sharing arguments for the employer size–wage effect. The employer size–wage effect survives all of these attempts to explain it away. The following section returns to the question of how we might try to estimate the true value of ε.

4.3 Reverse Regressions

The estimates so far suggest that, while an upward-sloping labor supply curve is a plausible explanation for part of the employer size–wage effect, the elasticity of wages with respect to employment is small (perhaps about 0.04) after we have controlled for observed and unobserved worker quality. Using the formula for the Hicks–Pigou rate of exploitation in (2.3) this elasticity is also the proportionate amount by which wages fall short of marginal product so an elasticity of 0.04 suggests only small deviations from perfect competition. But, there is no reason why one might not run a regression of log employment on the log wage (and the x variables), hope to estimate ε_{Nw} from the coefficient on wages in this regression, and then invert the elasticity to give ε. We will term this the reverse regression after the discussion in Goldberger (1984).

This regression provides an estimate of $E(\log(N_i) \mid \log(w_i), x_i)$. The following proposition tells us what we would expect to find in this case.

Proposition 4.2. *Running a regression of* $\log(N_i)$ *on* $(\log(w_i), x_i)$ *estimates*

$$E(\log(N_i) \mid \log(w_i), x_i) = \frac{1 - \rho'(\varepsilon + \eta)}{\varepsilon} \log(w_i) + \rho' \ln(1 + \varepsilon) - \frac{\beta_b - \rho'(\varepsilon\beta_a + \eta\beta_b)}{\varepsilon} x_i \tag{4.9}$$

where

$$\rho' \equiv \frac{\varepsilon\sigma_{ab} + \eta\sigma_b^2}{\varepsilon^2\sigma_a^2 + \eta^2\sigma_b^2 + 2\varepsilon\eta\sigma_{ab}} \tag{4.10}$$

Proof. See Appendix 4A.

[7] It is also worth noting that Teulings and Hartog (1998) find that the ESWE is smaller in "corporatist" countries where the level of wage bargaining is above that of the individual employer.

In the reverse regression, OLS only gives unbiased estimates of $(1/\varepsilon)$ if $\varepsilon\sigma_{ab} = -\eta\sigma_b^2$. But, of course, the bias will generally be different from that in the wage equation; compare (4.9) and (4.7). If $\sigma_{ab} = 0$, then the true estimate of ε must lie between that estimated by a regression of $\log(w)$ on $\log(N)$, and one obtained from a regression of $\log(N)$ on $\log(w)$.

Now compare the direct and reverse regressions. Table 4.5 presents some further estimates of (4.7) and (4.9) using data from the US CPS and the UK LFS that have already been used in table 4.2. In each row we report the coefficient on employer size when (4.7) is estimated in the column headed "coefficient on log employer size" while the inverse of the coefficient on the log wage when (4.9) is estimated is reported in the column headed "inverse of coefficient on log wage." As these are both estimates of ε, one would hope that these coefficients are similar.

In fact, they are very different. Consider the US results first. The first row reports the results when no other controls are included and subsequent rows add personal controls, education controls, regional controls, industry controls, and occupation controls, finishing with a model with all variables included. There are several general conclusions. First, for both the United States and the United Kingdom, there is always a large gap between the coefficient on log employer size and the inverse of the coefficient on the log wage. For example, in the first row of table 4.5, the second column suggests an elasticity of the wage with respect to employment of 0.116 while the third suggests an elasticity of 1.000. Secondly, the inclusion of controls always reduces the coefficient in the second column and raises the coefficient in the third column.

What is the explanation for this pattern of results? The gap in the two estimates of ε suggests the presence of the biases identified in (4.7) and (4.9) or of measurement errors in wages or employment, or both. But, why should this gap widen when controls are introduced?

As discussed earlier, the biases in (4.7) and (4.9) depend on relative variances and the correlation coefficient of the errors in the labor supply and MRPL equations, and in a complicated way. It is possible that the biases move in different directions when the introduction of controls changes the relative variances and the correlation coefficient but it is hard for intuition to deliver any firm expectation.

The measurement error argument of Griliches (1977) is perhaps more plausible so some consideration of the likely sources of measurement error in both employer size and wages is likely to be worthwhile. For employer size, individuals have no reason to know the size of their workplace and seem to often make big mistakes as the earlier discussion has made clear. If this is the case, one would expect the coefficient on a better measure of employer size in the direct regression to be larger. It is plausible to believe that managers in workplaces have better information than

TABLE 4.5
Direct and Reverse Regressions

Sample	*Coefficient on Log Employer Size*	*Inverse of Coefficient on Log Wage*	*Personal Controls*	*Education Controls*	*Regional Controls*	*Industry Controls*	*Occupation Controls*
US	0.116 (0.004)	1.000 (0.033)	No	No	No	No	No
US	0.099 (0.004)	1.070 (0.041)	Yes	No	No	No	No
US	0.071 (0.004)	1.273 (0.065)	Yes	Yes	No	No	No
US	0.095 (0.004)	1.110 (0.044)	Yes	No	Yes	No	No
US	0.085 (0.004)	1.392 (0.067)	Yes	No	No	Yes	No
US	0.078 (0.004)	1.181 (0.055)	Yes	No	No	No	Yes
US	0.060 (0.004)	1.608 (0.101)	Yes	Yes	Yes	Yes	Yes
UK	0.144 (0.003)	1.908 (0.037)	No	No	No	No	No
UK	0.115 (0.003)	1.988 (0.045)	Yes	No	No	No	No
UK	0.092 (0.002)	2.070 (0.054)	Yes	Yes	No	No	No
UK	0.110 (0.002)	1.964 (0.045)	Yes	No	Yes	No	No
UK	0.085 (0.003)	2.671 (0.088)	Yes	No	No	Yes	No
UK	0.075 (0.002)	2.368 (0.073)	Yes	No	No	No	Yes
UK	0.062 (0.002)	2.737 (0.109)	Yes	Yes	Yes	Yes	Yes

Notes.
1. US data are from the Contingent Worker Survey to the April 1993 CPS. The sample is restricted to the non-union sector. UK data are from the LFS.
2. The dependent variable is the log of the hourly wage.
3. Personal controls are sex, race, a quartic in experience, marital status, and the presence of children.
4. The column headed "coefficient on log employer size" gives the results from estimating a regression of log wages on log employer size and other controls. The column headed "inverse of coefficient on log wage" gives the results from estimating a regression of log employer size on log wages and other controls.

TABLE 4.6
The ESWE: Evidence from WERS

	Coefficient on Log Employer Size	*Inverse of Coefficient on Log Wage*	*Controls*	*Method of Estimation (Instrument)*
(1)	0.120 (0.010)	–	No	OLS
(2)	0.051 (0.007)	–	Yes	OLS
(3)	0.050 (0.007)	–	Yes	IV (first lag)
(4)	0.056 (0.007)	–	Yes	IV (fifth lag)
(5)	–	3.383 (0.365)	Yes	OLS
(6)	–	1.408 (0.188)	Yes	Between-firm

Notes.
1. Data are from the 1998 UK WERS. The number of observations is approximately 25,000. The dependent variable is the log of the hourly wage for rows (1)–(4) and the log of employers size in rows (5) and (6).
2. The controls included are sex, race, education, age, occupation, and industry. Observations are weighted to be representative of all workers in plants with more than 10 employees (smaller plants are excluded). Standard errors are computed assuming clustering on the workplace.

their workers about employer size, although it might be better to have administrative data like that used by Bayard and Troske (1999). The 1998 UK WERS can be used to investigate this. Some results are reported in table 4.6. The first row reports the result of a simple regression of log wages on log employer size. Log wages are reported by the individual workers concerned and employer size by the manager so these regressions are very similar to those reported earlier except that employer size is manager-reported. However, the coefficient on log employer size is similar to what we have seen before. The second row introduces controls: the drop in the coefficient on log employer size is very similar to what we have seen before. There is little evidence here to suggest substantial worker misreporting of employer size although the earlier results on the excessive apparent changes in employer size in the BHPS did suggest the existence of large measurement errors.[8]

An additional reason for why the coefficient on log employer size may be underestimated is the existence of transitory shocks to employment that are not reflected in wages. WERS also allows us a way to investigate this as employers are asked to report workplace employment one and five years previously. A simple regression of the log of employment this year on the log of employment last year has a coefficient of 0.97 on the lagged dependent variable suggesting a lot of permanence in the level of employment. The

[8] It is possible that the comparison of BHPS and WERS results is made difficult by the different nature of the two data sets.

third row of table 4.6 shows the coefficient on log(N) when it is instrumented by lagged employment to try to pick up the permanent component in employment. The instrumental variable coefficient is very similar to the ordinary least squares. The fourth row uses the fifth lag of employment as an instrument with very similar results. These results suggest that transitory shocks to employment that are not reflected in wages are not particularly important in explaining the estimated size of the ESWE.

The discussion so far has focused on measurement errors and transitory shocks to employment. But, there are reasons to think there might be similar problems surrounding wages. Perhaps the most important is that the wage used in the regressions in table 4.2 refers to the wage of a single worker in the plant. This is an unbiased estimate of the average log wage in the plant but obviously contains some measurement error. As a result, the coefficient on log wages in the reverse regression is likely to be biased towards zero. Evidence for this effect can be seen in the fifth and sixth rows of table 4.6. The fifth row reports the result of a reverse regression on the WERS data. The implied value of ε is very high. The sixth row reports the result of a between-plant regression on the same data, exploiting the fact that the data set contains observations on multiple workers within plants. This effectively runs a regression of the log of employment on the average log wage. The coefficient on the wage variable rises implying a lower value of ε, as one would expect.

None of the discussion so far gives us much confidence that OLS, either a direct or reverse regression, gives us a good estimate of ε. On the basis of the results reported, one could argue either that labor supply to the firm is very inelastic or that it is quite elastic. It is perhaps better to conclude that these regressions are just not very informative. But, while it might have been nice for OLS to deliver reliable results, perhaps it was expecting too much and a simpler potential solution presents itself.

To identify the labor supply curve (which is all we want here), a variable that shifts the MRPL curve without shifting the supply curve is needed. One can then use this as an instrument for the wage or employment in estimating the supply curve (depending on which way round we are estimating the supply curve). This procedure will yield consistent estimates of ε. But, of course, it requires us to be able to provide such an instrument.

If one is interested in estimating the elasticity of labor supply to an individual firm, then the instrument needs to be something that affects the demand curve for that firm but has negligible impact on the labor market as a whole. The reason is that a pervasive labor market demand shock will raise the general level of wages so is likely to be correlated with B in (4.2). So, for example, the approach of using demand shocks caused by exchange rate fluctuations (as in Abowd and Lemieux 1993) does not seem viable here.

There are a number of studies that attempt to use firm-level instruments. For example, Sullivan (1989) uses the population in the area surrounding the hospital as an instrument affecting the demand for nurses, and Beck et al. (1998) use the number of children in a school district as an instrument for the demand for teachers. These represent serious attempts to deal with a difficult problem but their instruments are not beyond criticism. If the main variation in the number of children or the number of patients comes from variation in population, it is also likely that the supply of nurses and teachers in an area is proportional to population as well. Perhaps the best studies are Staiger et al. (1999) and Falch (2001). Staiger et al. (1999) examine the impact of a legislated rise in the wages paid at Veteran Affairs hospitals. This combined with a plausible argument that these hospitals were allowed to hire as many staff as they wanted (which is required to make sure we are estimating the supply curve) seems as close to an exogenous increase in wages as anything else in the literature. They estimate the short-run elasticity in the labor supply to the firm to be very low—around 0.1 implying an enormous amount of monopsony power possessed by hospitals over their nurses. Falch (2001) investigates the impact on the supply of teachers to individual schools in Norway in response to a policy experiment that selectively raised wages in some schools. Again, he finds that the elasticity in the supply of labor to individual schools is very low. How plausible are these estimates and whether they can be generalized to the rest of the labor market are open questions as the markets for nurses and teachers are ones that might conventionally be thought of as having some monopsonistic elements.

This is all rather depressing: a good estimate of the elasticity of the labor supply curve facing the firm seems very elusive so perhaps there is a very good reason for the lack of research into this area. Progress seems to be dependent on finding a good firm-level instrument.

4.4 Estimating Models of Dynamic Monopsony

The previous part of this chapter has used the static model of monopsony as a way to think about the issue of estimating the elasticity of the labor supply curve facing an individual firm. The rest of this chapter uses a more explicitly dynamic, theoretical approach to estimate this elasticity. In a steady state, we know that the supply of labor to the firm $N(w)$ must be given by $N(w) = R(w)/s(w)$ where $R(w)$ is the flow of recruits to the firm and $s(w)$ is the separation rate. As pointed out by Card and Krueger (1995) and discussed earlier in section 2.2, this implies that

$$\varepsilon_{Nw} = \varepsilon_{Rw} - \varepsilon_{sw} \tag{4.11}$$

so that knowledge of the elasticities of recruitment and quits with respect to the wage can be used to estimate the elasticity of labor supply facing the firm. This section discusses how we can estimate ε_{Rw} and ε_{sw}.

One of the advantages of this approach is that there is a well-established literature that discusses the elasticity of the separation rate with respect to the wage (e.g., Pencavel 1972; Parsons 1972, 1973; Viscusi 1980; Light and Ureta 1992). However, an apparent disadvantage is that it might be unclear how the elasticity of the recruits with respect to the wage should be estimated. Card and Krueger (1995) use estimates of the elasticity of job applicants with respect to the wage from Holzer et al. (1991) and Krueger (1988) but the justification for this is not obvious. One of the contributions here is to show that there is a close connection between ε_{Rw} and ε_{sw} so that estimates of the separation elasticity are informative about the recruitment elasticity.

Consider the basic Burdett and Mortensen (1998) model of dynamic monopsony introduced in section 2.4. In this model, we have

$$s(w) = \delta + \lambda[1 - F(w)] \tag{4.12}$$

$$R(w) = R^{u} + \lambda \int^{w} f(x)N(x)dx \tag{4.13}$$

where $s(w)$ is the separation rate in a firm that pays wage w, $R(w)$ is the flow of recruits, δ is the rate (assumed exogenous) at which workers leave employment for non-employment, λ is the arrival rate of job offers, $F(w)$ is the distribution of wage offers, R^{u} are the recruits from unemployment (which does not depend on the wage offered) and $N(w)$ is the employment level in a firm that pays w. By differentiating (4.12) and (4.13), we have

$$\varepsilon_{sw} = \frac{ws'(w)}{s(w)} = -\frac{\lambda w f(w)}{s(w)} = -\frac{\lambda w f(w)N(w)}{R(w)} = -\frac{wR'(w)}{R(w)} = -\varepsilon_{Rw} \tag{4.14}$$

where the third equality sign follows from the fact that, in steady state, $s(w)N(w) = R(w)$. (4.14) says that, in a steady state, the recruitment elasticity is simply minus the separation elasticity so that (using (4.11)) one can simply double the separation elasticity to get an estimate of the labor supply elasticity. The explanation for the connection between the two elasticities is that separations from one firm for a wage-related reason must be the recruit of some other firm so that quits and recruits are two sides of the same coin.

Although this result is neat, one might wonder about its robustness. So, let us consider some generalizations. First, relax the assumption that workers always quit for a better-paying job and never quit for a job with lower pay. Suppose that a worker currently being paid w accepts

a job offer of x with probability $\phi(x/w)$. We assume that it is the ratio of the wages that matters which seems a reasonable restriction as there is no reason to think that a general increase in wages would have any effect on job-to-job mobility rates. Note that the model of (4.12) and (4.13) corresponds to the case where $\phi(x/w) = 0$ if $x < w$ and $\phi(x/w) = 1$ if $x > w$. The separation rate and recruitment functions will now be given by

$$s(w) = \delta + \lambda \int \phi\left(\frac{x}{w}\right) f(x)dx \tag{4.15}$$

$$R(w) = R^u + \lambda \int \phi\left(\frac{w}{x}\right) f(x)N(x)dx \tag{4.16}$$

The following proposition tells us that a suitably weighted separation elasticity must be equal to a suitably weighted recruitment elasticity.

Proposition 4.3. *If the separation and recruitment functions are given by (4.15) and (4.16), then the recruit-weighted separation and recruitment elasticities must be equal, that is,*

$$\int \varepsilon_{sw}(w)R(w)f(w)dw = -\int \varepsilon_{Rw}(w)R(w)f(w)dw \tag{4.17}$$

Proof. See Appendix 4A.

If the separation and recruitment elasticities are both constant, (4.17) says that they must be equal. If they vary with the wage then they can differ but probably not by much. For example, if the separation elasticity is finite everywhere, the recruitment elasticity cannot be infinite for any positive measure of employees.

However, the separation and recruitment functions of (4.15) and (4.16) are still quite restrictive in that they assume that separations to and recruitment from non-employment are not sensitive to the wage.

Write total separations as $s(w) = s^n(w) + s^e(w)$ where $s^n(w)$ is the separation rate to non-employment and $s^e(w)$ is the separation rate to employment. Denote by θ_s the share of separations which are a direct move to another job. Similarly write total recruits as $R(w) = R^n(w) + R^e(w)$ where $R^n(w)$ is the flow of recruits from non-employment and $R^e(w)$ is the flow of recruits from employment. Denote by θ_R the share of recruits from employment. The overall elasticity of labor supply with respect to the wage can be written as

$$\varepsilon_{Nw} = \theta_R \varepsilon^e_{Rw} + (1 - \theta_R)\varepsilon^n_{Rw} - \theta_s \varepsilon^e_{sw} - (1 - \theta_s)\varepsilon^n_{sw} \tag{4.18}$$

so that knowledge of the four elasticities can be used to compute the

elasticity of the labor supply curve facing the firm. The two separation elasticities can be estimated straightforwardly (see below) so the only problem is how to estimate the recruitment elasticities. The following proposition shows that a suitably weighted recruitment elasticity from employment is equal to a weighted separation elasticity to employment.

Proposition 4.4. *If the separation and recruitment functions to and from employment are given by*

$$s^e(w) = \lambda \int \phi\left(\frac{x}{w}\right) f(x)dx \tag{4.19}$$

$$R^e(w) = \lambda \int \phi\left(\frac{w}{x}\right) f(x)N(x)dx \tag{4.20}$$

then a suitably weighted average of the separations elasticity must be equal to a weighted average of the recruitment elasticities. In particular,

$$\frac{\int \varepsilon^e_{sw}(w)s^e(w)N(w)f(w)dw}{\int s^e(w)N(w)f(w)dw} = -\frac{\int \varepsilon^e_{Rw}(w)R^e(w)f(w)dw}{\int R^e(w)f(w)dw} \tag{4.21}$$

Proof. See Appendix 4A.

(4.21) says that a weighted average of the separation elasticity and the recruitment elasticity must be equal but that the weights are not equal, as in (4.17).[9] However, this still means that if both the separation and recruitment elasticities are constant, they must be equal.

Unfortunately, there is no equivalent result relating the separation and recruitment elasticities for transitions to and from non-employment. For example, if there is heterogeneity in the reservation wages of workers but the reservation wage for an individual worker never changes, recruits from non-employment are increasing in the wage but separations to non-employment are not. However, if there is a stochastic component to the reservation wage, separations to non-employment are also sensitive to the wage. The basic problem is that, whereas a move from one job to another is a quit for one firm and immediately a recruit for another firm, this is not true of flows between employment and non-employment.

So, we cannot use the wage elasticity of separations to non-employment to estimate the wage elasticity of recruits from non-employment: we need a different method. The share of recruits from employment is given by

[9] The separation elasticity is weighted by separations to other jobs while the recruitment elasticity is weighted by recruits from other jobs. We would expect the weight on the elasticity in high-wage firms to be larger for the recruitment than the separation elasticity.

$$\theta_R(w) = \frac{R^e(w)}{R^e(w) + R^n(w)} \tag{4.22}$$

This enables us to prove the following relationship between $\varepsilon^e_R(w)$ and $\varepsilon^n_R(w)$.

Proposition 4.5. *If the share of recruits from employment is given by (4.22), then the relationship between the wage elasticity of recruits from non-employment, $\varepsilon^n_{Rw}(w)$, and the wage elasticity of recruits from employment, $\varepsilon^e_{Rw}(w)$, is given by*

$$\varepsilon^n_{Rw}(w) = \varepsilon^e_{Rw}(w) - \frac{w\theta'_R(w)}{\theta_R(w)[1 - \theta_R(w)]} \tag{4.23}$$

Proof. See Appendix 4A.

If, for example, we model $\theta_R(w)$ as a logistic function $e^{\beta x}/(1 + e^{\beta x})$ where x includes the log wage, then $(w\theta'_R/\theta_R(1 - \theta_R)) = \beta_w$ where β_w is the coefficient on the log wage.

Summarizing all this information, our strategy for estimating the elasticity of the labor supply curve facing the firm is

- estimate separations equations for separations to employment and non-employment, and obtain the wage elasticities;
- use the wage elasticity of separations to employment to estimate the wage elasticity of recruits from employment (based on Proposition 4.4);
- estimate a logit model for the probability that a recruit comes from employment and then use (4.23) to estimate the elasticity of recruits from non-employment;
- use these elasticities and information of the share of separations to and recruits from employment in (4.14) to estimate the elasticity of the labor supply curve facing the firm.

Let us now put this into practice.

4.5 Estimating the Wage Elasticity of Separations

We model the instantaneous separation rate as $s = e^{\beta x}$. One of the x variables will be the log of the wage so that the elasticity of the separation rate with respect to the wage will simply be the coefficient on the wage. From the previous discussion, we also need to model the separations to employment and non-employment separately. Write the separation rate to other jobs as $s^{ee}(x) = \exp(\beta^{ee}x)$ and the separation rate to non-employ-

ment as $s^{en}(x) = \exp(\beta^{en}x)$. We assume that, conditional on x, the two sorts of separation are independent.

Define an indicator variable y^{en} which takes the value 1 if the individual has a spell of non-employment in a period of time τ and 0 otherwise and an another indicator variable y^{ee} which, if the individual does not have a spell of non-employment, takes the value 1 if the individual changes jobs and 0 if they do not. The probabilities of the different outcomes are given by

$$\Pr(y^{en} = 1 \mid x) = 1 - \exp(-s^{en}(x)\tau)$$

$$\Pr(y^{en} = 0, y^{ee} = 1 \mid x) = \exp(-s^{en}(x)\tau)(1 - \exp(-s^{ee}(x)\tau)) \quad (4.24)$$

$$\Pr(y^{en} = 0, y^{ee} = 0 \mid x) = \exp(-s^{en}(x)\tau)\exp(-s^{ee}(x)\tau)$$

so that the individual contribution to the log-likelihood function can be written as

$$\log L = y^{en} \ln[1 - \exp(-s^{en}(x)\tau)] + (1 - y^{en}) \ln[\exp(-s^{en}(x)\tau)]$$

$$+(1 - y^{en})[y^{ee} \ln[1 - \exp(-s^{ee}(x)\tau)] + (1 - y^{ee}) \ln[\exp(-s^{ee}(x)\tau)]] \quad (4.25)$$

The important feature of (4.25) is that one can estimate the separations elasticity to non-employment and other jobs separately. To estimate the elasticity of separations to non-employment, the whole sample is used and we have as a dependent variable whether the individual had a period of non-employment in the year. To estimate the elasticity of separations to other jobs the sample of those who have been in continuous employment is used and we have as a dependent variable whether the individual remains in the same job. Note that the overall elasticity will be a weighted average of these two elasticities, the weight being the fraction of separations that are to non-employment.

The most serious problem in estimating the wage elasticities is, as always, going to be the result of a failure to control adequately for other relevant factors. One potential source of problems in estimating the separation elasticity is a failure to control adequately for the average level of wages in the individual's labor market. Separations are likely to depend on the wage relative to this alternative wage so that a failure to control for the alternative wage is likely to lead to a downward bias in the wage elasticities. On the other hand, we would expect separations to be more sensitive to the permanent component of wages than to the part of wages that is a transitory shock or measurement error. In this case, the inclusion of controls correlated with the permanent wage is likely to reduce the estimated wage elasticity. Table 4.7 estimates some separations equations with and without controls for the PSID, NLSY, BHPS and

TABLE 4.7
The Sensitivity of the Separation Elasticity to Specification

	PSID (US)	NLSY (US)	BHPS (UK)	LFS (UK)
All separations				
Mean separation rate	0.21	0.55	0.19	0.058
No controls	−0.944 (0.030)	−0.515 (0.019)	−0.798 (0.032)	−0.646 (0.021)
With controls	−0.973 (0.041)	−0.536 (0.032)	−0.720 (0.041)	−0.500 (0.028)
Tenure controls	−0.575 (0.037)	−0.340 (0.026)	−0.503 (0.064)	−0.343 (0.032)
Separations to employment				
Mean separation rate	0.12	0.43	0.12	0.032
No controls	−0.759 (0.050)	−0.307 (0.018)	−0.631 (0.038)	−0.529 (0.030)
With controls	−0.867 (0.038)	−0.359 (0.032)	−0.688 (0.049)	−0.425 (0.039)
Tenure controls	−0.450 (0.042)	−0.156 (0.027)	−0.429 (0.050)	−0.207 (0.044)
Separations to non-employment				
Mean separation rate	0.08	0.12	0.07	0.025
No controls	−1.010 (0.067)	−0.750 (0.028)	−0.916 (0.048)	−0.748 (0.029)
With controls	−0.892 (0.087)	−0.850 (0.055)	−0.632 (0.066)	−0.578 (0.041)
Tenure controls	−0.569 (0.068)	−0.713 (0.056)	−0.493 (0.071)	−0.477 (0.045)

Notes.
1. This table reports the elasticities of separations with respect to the wage. The PSID, NLSY, and BHPS samples are those described in the Data Sets Appendix. The LFS sample is from September 1997 to November 1999. The LFS also differs from the other data sets in modeling labor market transitions from one quarter to another instead of one year to another. This is why the means of the dependent variables are so much lower. The row headed "no controls" simply includes the wage. The rows marked "with controls" include gender, education, race, marital status, children, region, a quartic in experience, and year dummies. The row headed "tenure controls" includes a quartic in tenure in addition to the usual controls.

LFS. First consider the wage elasticity for all separations (the top panel of table 4.7). All the estimated wage elasticities are negative and significantly different from zero. They range from −0.5 for the NLSY to −0.9 for the PSID. The bottom two panels estimate separate wage elasticities for separations to employment and non-employment. There is evidence that the elasticities for both separations to employment and non-employment are both sensitive to the wage with the latter being larger than the former.

Inclusion of standard human capital controls does not make much difference to the estimated wage elasticities. However, one variable whose inclusion or exclusion makes a lot of difference to the apparent estimated wage elasticity is job tenure.[10] The inclusion of job tenure always drastically reduces the estimated wage elasticity as high-tenure workers are less likely to leave the firm and are more likely to have high wages. There are arguments both for and against the inclusion of job tenure. One of the benefits of paying high wages is that tenure will be higher so that one needs to take account of this indirect effect if one wants the overall wage elasticity when including tenure controls: in this situation, excluding tenure may give better estimates. On the other hand, if there are seniority wage scales, the apparent relationship between separations and wages may be spurious.

Unobserved heterogeneity that is correlated with the wage causes familiar problems but the wage elasticity is likely to be biased even if there is heterogeneity uncorrelated with the wage. To see this, suppose that the separation rate is $\xi w^{-\beta}$ where ξ is unobserved and independent of the wage. To keep things tractable we will assume that ξ has a gamma distribution with mean μ and variance σ^2. The following proposition summarizes the effect of unobserved heterogeneity on the estimated wage elasticity.

Proposition 4.6. *The elasticity of the separations rate with respect to the wage is biased towards zero with gamma-distributed unobserved heterogeneity that is uncorrelated with the wage. The shorter the time period over which the data are observed, the smaller is the bias.*

Proof. See Appendix 4A.

The result on the existence of a bias is unsurprising: we are using survivor functions to estimate the wage elasticity and it is well known that unobserved heterogeneity has an effect on the estimated coefficients in duration models (see Lancaster 1990). Table 4.8 investigates whether the time horizon makes any difference to the estimated wage elasticity

[10] The word "apparent" is appropriate here because the dependence of job tenure on the wage needs to be taken into account when estimating the full wage elasticity.

TABLE 4.8
The Effect of the Time Horizon on the Separation Elasticity: UK LFS

Controls	*No*	*Yes*
1 quarter	−0.646 (0.021)	−0.500 (0.028)
2 quarters	−0.640 (0.018)	−0.497 (0.024)
3 quarters	−0.586 (0.017)	−0.471 (0.023)
4 quarters	−0.547 (0.017)	−0.429 (0.023)

Notes.
1. The reported coefficients are the coefficients on the log wage from the estimated separations model as described in section 4.5. The controls included are gender, race, education, experience, martial status and dependent children, region, and month.
2. The dependent variable in the row marked 1 quarter is whether the individual left the initial job over the first quarter, that for 2 quarters whether the individual left over the first two quarters, etc.

using data from the UK LFS. Results are reported for the wage elasticity estimated over a period of one to four quarters. The results are consistent with the predictions of Proposition 4.6. Both with and without other controls, the estimated wage elasticity is higher over short periods than long. As all the other data sets used in table 4.3 use a time horizon of a year, this suggests the estimates may be understating the true wage elasticity. However, the results in table 4.8 do not suggest the size of the bias is large. The estimated elasticity of separations with respect to the wage rises (in absolute terms) from −0.43 to −0.5 as the time horizon is narrowed from one year to one quarter and hardly narrows at all as the time horizon falls from two quarters to one quarter.

4.6 The Proportion of Recruits from Employment

Proposition 4.5 says that we need to know how the fraction of recruits from employment varies with the wage. Table 4.9 reports the results of estimating a logit model for a recruit coming from employment for our four data sets. In all data sets, the higher the wage, the higher the probability that a recruit comes from employment: this is as the theory predicts. This implies that the wage elasticity of recruits from employment is higher than the wage elasticity of recruits from non-employment.

4.7 The Elasticity of the Labor Supply Curve Facing the Firm

We are now in a position to provide an estimate of the labor supply curve facing the firm using the results of the previous two sections and (4.18).

TABLE 4.9
The Probability of a Recruit Coming from Employment

Data Set	*Mean of Dependent Variable*	*Coefficient (SE) on Log Wage Without Controls*	*Coefficient (SE) on Log Wage with Controls*	*Number of Observations*
PSID	0.29	1.011 (0.036)	0.948 (0.054)	14277
NLSY	0.32	0.533 (0.037)	0.674 (0.042)	13653
BHPS	0.36	1.129 (0.065)	1.384 (0.080)	4649
LFS	0.51	0.824 (0.035)	0.746 (0.042)	12071

Notes.
1. The dependent variable is a dummy variable taking the value 1 if a worker was recruited from employment and 0 otherwise. The sample is all recruits. The estimation method is logit. The other controls are gender, race, experience, education, region, and year dummies.

The results of this are reported in table 4.10 where no attempt has been made to correct the wage elasticities for the problems caused by measurement error and the time horizon. These labor supply elasticities are low—in the region of 1. There are a number of reasons why one might argue that these elasticities are underestimates but it is clear that extremely large adjustments to these estimates are necessary to make perfect competition an acceptable approximation. For example, the evidence on the size of the bias caused by the interaction of the time horizon and unobserved heterogeneity would not dramatically increase these elasticities. One can only conclude that the elasticity of separations with respect to the wage is low and that this results in a low elasticity in the supply of labor to individual employers.

TABLE 4.10
The Elasticity of the Labor Supply Curve Facing the Firm

Data	*PSID*	*NLSY*	*BHPS*	*LFS*
Elasticity of separations to employment	0.867	0.359	0.631	0.529
Elasticity of separations to non-employment	0.892	0.850	0.632	0.578
Share of separations to employment	0.62	0.78	0.63	0.56
β_w	0.948	0.674	1.384	0.746
Elasticity of labor supply curve	1.38	0.68	0.75	0.75

Notes.
1. These computations used table 4.7 for the separation elasticities and the share of separations to employment, table 4.9 for the estimate of β_w, and (4.18) for the elasticity of the labor supply curve. The share of recruits from employment is assumed to be equal to the share of separations to employment as must be the case in steady state.

4.8 The Estimation of Structural Equilibrium Search Models of the Labor Market

This is the best place in this book to discuss a small but relevant literature on the structural estimation of equilibrium search models. The earliest empirical model, Eckstein and Wolpin (1990), used the Albrecht and Axell (1984) model but most later contributions have based their empirical analysis on some variant of the Burdett and Mortensen (1998) model described in section 2.4 (for surveys, see van den Berg 1999; Mortensen 2002). Because the parameters of that model contain all the necessary information (assuming, of course, that the model is correctly specified) for working out the supply of labor facing the firm, these estimates contain within them an estimate of the wage elasticity of the labor supply to an individual firm. Sometimes this elasticity is made explicit, although more often it is not.

Whether this general equilibrium approach to empirical modeling is a superior methodology depends on the purpose to which one is going to put the estimates. If one wants to model the general equilibrium effects of a change in an economy-wide policy, then such an approach may be essential. But if one simply wants an estimate of the extent of employer market power there are reasons to think that the structural approach offers few advantages and many disadvantages. The earliest estimates of equilibrium search models (e.g., Kiefer and Neumann 1993; van den Berg and Ridder 1998) emphasized how the general equilibrium model provided a tight link between the wage distribution and labor market transition rates that was exploited in the structural estimation. But, this link was more of a problem than a help and the models did not explain the distribution of wages well. More sophisticated models were introduced that added employer and worker heterogeneity (see Bontemps et al. 1999, 2000) and generalizing the wage policy used by employers (see Postel-Vinay and Robin 2002). The effect of these reasonable generalizations is effectively to allow the distribution of wages to vary independently of the labor market frictions. In this case, joint estimation of transition rates and the wage distribution offers no real advantages, and the complication of the models hides the simple economics at work. For all their sophistication one ends up with estimates of the extent of frictions that are not much more advanced than the back-of-the-envelope calculations in section 2.6. And, the partial equilibrium approach described in this chapter dominates that approach as it makes less in the way of stringent assumptions about the economy. So, it is probably the case that these equilibrium search models have told us little about the extent of frictions in the labor market that we could not have learned in their absence.

The discussion here mirrors a wider debate in econometrics about the merits of structural modeling. Structural models provide excellent estimates if the model is correct but may not be robust to small deviations from the maintained model, and tractability may restrict attention to implausibly simplistic models. A more pragmatic approach is likely to provide more robust estimates that may not be fully efficient if one pretends to know the true model. As Wolpin (1992: 558) put it 10 years ago "methods for estimating dynamic stochastic models of this kind are still in a relatively undeveloped stage, and knowledge about the effects of model and solution misspecification is very limited. ... exactly, how seriously one should take these particular estimates as reflecting real phenomena is open to debate." Unfortunately, 10 years on, the debate remains just as open and little progress has been made. For example, a state-of-the-art paper in the area (Postel-Vinay and Robin 2002) apparently shows that a model based on the totally implausible assumption of universal offer-matching (with the implication, among others, that unemployed workers are indifferent about getting a job or not) to maximize profits (in an economy, France, where union coverage approaches 100%) is consistent with the observed data, it is time to start worrying about identification as at least one other model of the labor market (the correct one) must also be consistent with the data. My gut feeling is that, for the purpose of estimating the wage elasticity in the supply of labor to an individual employer, a pragmatic approach is more likely to deliver credible results.

4.9 Conclusions

The fundamental difference between monopsony or oligopsony and perfect competition is the size of the elasticity of the labor supply curve facing a firm. Perfect competition assumes it is infinite, imperfect competition that it is finite. There is remarkably little literature on estimating the elasticity of this labor supply curve and this chapter has tried to fill that gap. It has investigated two main methods: one based on the correlation between wages and employment (the employer size–wage effect) and the other based on the estimation of separations functions. The two approaches give rather different results. OLS estimates of the ESWE suggest that the wage elasticity of the labor supply curve to individual employers is in the region of 10–15 leading to a wage that is 6–10% below marginal revenue product. However, there are reasons to think that OLS overstates the wage elasticity and IV estimates are a lot lower. The approach based on estimating the wage elasticity of separations suggests a wage elasticity of the labor supply curve to individual

employers that is around one. While there are reasons to think these may be underestimates, it would certainly be hard to argue on the basis of these estimates that the labor supply elasticity is anywhere near 10. However, neither approach is entirely satisfactory: progress really needs good firm-level instruments although these are likely to be hard to find. Given this, it would probably be unwise to base one's belief in the market power of employers too much on these estimates if there were no other evidence. But, as the rest of the book sets out to show, there are very good reasons for believing that employers do have non-negligible market power.

Appendix 4A

Proof of Proposition 4.1

Taking logs of (4.2), we have

$$E(\log(w_i) \mid \log(N_i), x_i) = E(b_i \mid \log(N_i), x_i) + \varepsilon \log(N_i)$$

$$= \beta_b x_i + E(\nu_{bi} \mid \log(N_i), x_i) + \varepsilon \log(N_i) \quad (4.26)$$

Now, from (4.4) and (4.6) we can derive

$$E(\nu_{bi} \mid \log(N_i), x_i)$$

$$= E(\nu_{bi} \mid \nu_{ai} - \nu_{bi} = (\varepsilon + \eta)\log(N_i) - (\beta_a - \beta_b)x_i + \ln(1 + \varepsilon))$$

$$= \frac{\sigma_{ab} - \sigma_b^2}{\sigma_a^2 + \sigma_b^2 - 2\sigma_{ab}}[(\varepsilon + \eta)\log(N_i) - (\beta_a - \beta_b)x_i + \ln(1 + \varepsilon)] \quad (4.27)$$

where the second equality follows from standard results on bivariate normal distributions. This can be written as (4.7) and (4.8).

Proof of Proposition 4.2

Taking the log of (4.2), leads to

$$E(\log(N_i) \mid \log(w_i), x_i) = \frac{1}{\varepsilon}\log(w_i) - \frac{1}{\varepsilon}E(b_i \mid \log(w_i), x_i)$$

$$= \frac{1}{\varepsilon}\log(w_i) - \frac{\beta_b}{\varepsilon}x_i - \frac{1}{\varepsilon}E(\nu_{bi} \mid \log(w_i), x_i) \quad (4.28)$$

Now, from (4.5) and (4.6) and standard results on the bivariate normal distribution, we can derive

$$E(\nu_{bi} \mid \log(w_i), x_i)$$

$$= E(\nu_{bi} \mid \varepsilon\nu_{ai} + \eta\nu_{bi} = (\varepsilon + \eta)\log(w_i) - (\varepsilon\beta_a + \eta\beta_b)x_i + \varepsilon\ln(1 + \varepsilon))$$

$$= \frac{\varepsilon\sigma_{ab} + \eta\sigma_b^2}{\varepsilon^2\sigma_a^2 + \eta^2\sigma_b^2 + 2\varepsilon\eta\sigma_{ab}}[(\varepsilon + \eta)\log(w_i) - (\varepsilon\beta_a + \eta\beta_b)x_i + \varepsilon\ln(1 + \varepsilon)] \tag{4.29}$$

which, substituting into (4.28) leads to (4.9) and (4.10).

Proof of Proposition 4.3

Differentiating (4.15), we have

$$s'(w) = -\lambda \int \frac{x}{w^2}\phi'\left(\frac{x}{w}\right) f(x)dx \tag{4.30}$$

so that

$$\int \varepsilon_{sw}(w)R(w)f(w)dw = \int \frac{ws'(w)}{s(w)}R(w)f(w)$$

$$= -\lambda \int\int \frac{x}{w}\phi'\left(\frac{x}{w}\right) f(x)N(w)f(w)dxdw \tag{4.31}$$

where we have used the steady-state relation $sN = R$. Now, exchanging the roles of x and w in (4.16) and differentiating with respect to x we have

$$R'(x) = \lambda \int \frac{1}{w}\phi'\left(\frac{x}{w}\right) f(w)N(w)dw \tag{4.32}$$

so that (4.31) can be written as

$$\int \varepsilon_{sw}(w)R(w)f(w)dw = -\int xR'(x)f(x)dx = -\int \varepsilon_{Rw}(x)R(x)f(x)dx \tag{4.33}$$

which is (4.17).

Proof of Proposition 4.4

Differentiating (4.19) with respect to w, we have

$$s^{\epsilon'}(w) = -\lambda \int \frac{x}{w^2}\phi'\left(\frac{x}{w}\right) f(x)dx \tag{4.34}$$

so that

$$\int \varepsilon^{e}_{sw}(w)s^{e}(w)N(w)f(w)dw = -\lambda \int\int \frac{x}{w}\phi'\left(\frac{x}{w}\right)f(x)N(w)f(w)dxdw \tag{4.35}$$

Exchanging the roles of x and w in (4.20) and differentiating with respect to x, we have

$$R^{e'}(x) = \lambda \int \frac{1}{w}\phi'\left(\frac{x}{w}\right)f(w)N(w)dw \tag{4.36}$$

so that (4.35) can be written as

$$\int \varepsilon^{e}_{sw}(w)s^{e}(w)N(w)f(w)dw = -\int xR^{e'}(x)f(x)dx = -\int \varepsilon^{e}_{Rw}(x)R^{e}(x)f(x)dx \tag{4.37}$$

Now, in the economy as a whole (but not firm by firm), total recruits from employment must equal total separations to employment which means that $\int s^{e}(w)N(w)f(w)dw = \int R^{e}(w)f(w)dw$. Dividing both sides of (4.37) by this leads to (4.21).

Proof of Proposition 4.5

Rearranging (4.22), we have

$$R^{n} = \frac{1-\theta_R}{\theta_R}R^{e} \tag{4.38}$$

which can be written as

$$\log(R^{n}) = \log(R^{e}) + \log\left(\frac{1-\theta_R}{\theta_R}\right) \tag{4.39}$$

Differentiation leads to (4.23).

Proof of Proposition 4.6

Denote the density function of ξ by $\varphi(\xi)$. The survivor function over a length of time τ will now by given by

$$S(w,\tau) = \int \exp(-\xi_w - \beta_\tau)\varphi(\xi)d\xi \tag{4.40}$$

and the estimate of the wage elasticity will be given by the elasticity of $-\log(S(w,\tau))$ with respect to the wage.

If ξ has a gamma distribution, then (4.40) can be written as

$$S(w,\tau) = \frac{\mu/\sigma^2}{\Gamma(\theta)}\int \exp(-\xi_w - \beta_\tau)\exp\left(-\xi\frac{\mu}{\sigma^2}\right)\left(\frac{\mu\xi}{\sigma^2}\right)^{(\mu/\sigma)^2}d\xi \tag{4.41}$$

After some rearrangement, this can be written as

$$S(w,\tau) = \left(\frac{(\mu/\sigma^2)}{w^{-\beta\tau} + (\mu/\sigma^2)}\right)^{\xi} \frac{1}{\Gamma(\xi)} \int \exp\left(-\xi_w - \beta_\tau + \frac{\mu}{\sigma^2}\right)$$
$$\times \left(w^{-\beta_\tau} + \left(\frac{\mu}{\sigma^2}\right)\right)^{(\mu/\sigma)^2} \xi^{(\mu/\sigma)^2 - 1} d\xi$$
$$= \left(\frac{(\mu/\sigma^2)}{w^{-\beta_\tau} + (\mu/\sigma^2)}\right)^{(\mu/\sigma)^2} \tag{4.42}$$

Taking logs, we have

$$(w,\tau) = \left(\frac{\mu}{\sigma}\right)^2 \left(\log(\mu) - \log(\mu + \sigma^2 w^{-\beta_\tau})\right) \tag{4.43}$$

Taking the elasticity of $-\log(S)$ with respect to the wage leads to

$$\frac{\partial \log(-(w,\tau))}{\partial \log(w)} = -\beta \frac{(\sigma^2/\mu)w^{-\beta_\tau}}{1 + (\sigma^2/\mu)w^{-\beta_\tau}} \frac{1}{\log(1 + (\sigma^2/\mu)w^{-\beta_\tau})} > -\beta \tag{4.44}$$

(4.44) shows that the estimated wage elasticity is biased towards zero by the presence of unobserved heterogeneity. The size of the bias is increasing in τ so the bias will be lower when a shorter period is used for estimation.

Appendix 4B

This appendix considers two generalizations of the static model of section 4.1.

The Employer Size–Wage Effect and Dynamic Labor Supply Curves

For the most part, our regressions have been of the current wage on current employment. As there is likely to be a difference between the short-run and long-run elasticity of the labor supply curve (see the discussion in section 2.2), one might wonder which elasticity is estimated using cross-sectional data when the true labor supply curve is dynamic.

Suppose the dynamic labor supply curve can be written in the following log-linear isoelastic form:[11]

[11] This might come from the equation $N_t - N_{t-1} = R(w_t) - s(w_t)N_{t-1}$ which says that the change in employment is the difference between recruits and quits.

$$w_t = \varepsilon^s(n_t - n_{t-1}) + \varepsilon n_{t-1} + \nu_{wt} \tag{4.45}$$

where ε^s is the short-run elasticity and ε the long-run elasticity. When one estimates a static regression of w_t on n_t one will estimate

$$E(w_t \mid n_t) = \varepsilon^s n_t + E((\varepsilon - \varepsilon^s)n_{t-1} + \nu_{wt} \mid n_t) \tag{4.46}$$

To work out the last term one needs to know the correlation between n_t and n_{t-1}. A simple model is the following:

$$n_t = \beta n_{t-1} + \nu_{nt} \tag{4.47}$$

where β is a measure of the persistence in employment. One should think of this as being a reduced-form equation for employment. We will assume that $\nu_t = (\nu_{wt}, \nu_{nt})$ is independent of n_{t-1} and jointly normally distributed with mean zero and covariance matrix Σ. Denote by σ_w^2 the variance of ν_w, σ_n^2 the variance of ν_n and σ_{wn} the covariance between ν_w and ν_n. Given these assumptions, the unconditional distribution of n_t (and n_{t-1}) will be normal with variance $\sigma_n^2/(1 - \beta^2)$ and $\mathrm{Cov}(n_t, \nu_{nt}) = \sigma_n^2$. Hence, we will have

$$E(\nu_{wt} \mid \nu_{nt}) = \frac{\sigma_{wn}}{\sigma_n^2}\nu_{nt}$$

$$E(\nu_{nt} \mid n_t) = (1 - \beta^2)n_t \tag{4.48}$$

Putting these into (4.48) leads to

$$E(w_t \mid n_t) = \left(\beta\varepsilon + (1 - \beta)\varepsilon^s + \frac{\sigma_{wn}(1 - \beta^2)}{\sigma_n^2}\right)n_t \tag{4.49}$$

The last term is the simultaneous equations bias caused by the potential correlation between the errors in wage and employment equation: this term could be eliminated by the use of suitable instruments. The other term shows that the estimated elasticity will be a weighted average of the short- and long-run elasticities with the weight being determined by the persistence in employment. So, if employment has no persistence, we will estimate the short-run supply curve and if it has full hysteresis, then we will estimate the long-run elasticity. As the evolution of employment within plants seems quite close to a random walk, it is likely that the cross-sectional correlation between wages and employment estimates the long-run elasticity.

The Employer Size–Wage Effect and the Labor Cost Function

Section 2.3 introduced the generalized model of monopsony and recommended the use of the labor cost function to think about the extent of

monopsony in the labor market. Yet, section 4.1 has reverted to a simple monopsony model in which the wage is the only instrument available to the employer for influencing its supply of labor. In this section, we show that the conclusions of the previous section are robust to using the labor cost function approach. Recall that the labor cost function $C(w, N)$ gave the per worker costs of recruitment and training if the firm pays a wage w and wants to have employment of N. To capture this idea assume that, if the firm spends C per worker on recruitment/training activities, its labor supply curve, (4.2), is modified to become

$$w = BC^{-\gamma}N^{\varepsilon} \tag{4.50}$$

where the isoelastic functional form is chosen for convenience. The formula for the rate of exploitation needs to be modified for the presence of C: the natural measure to use is $[Y' - w - C]/[w + C]$ as workers should not expect to receive their costs of training and recruitment.

As C is likely to be unobserved by the econometrician, one might think that the presence of C makes it very difficult to estimate the rate of exploitation. However, the following proposition shows that, once one has appropriately modified the formula for the rate of exploitation, the unobservability of C causes no problems and an estimate of the ESWE gives us the correct parameter estimate.

Proposition 4.8. *The rate of exploitation is given by*

$$\frac{Y'(N) - (w + C)}{w + C} = \frac{\varepsilon}{1 + \gamma} \tag{4.51}$$

and the "reduced-form" labor supply curve after concentrating out the optimal choice of C is given by

$$w = \gamma^{-\gamma/(1+\gamma)} B^{1/(1+\gamma)} N^{\varepsilon/(1+\gamma)} \tag{4.52}$$

Proof. Given N, C will be chosen to minimize $(w + C)$ which, using (4.50), leads to the first-order condition

$$1 = \gamma B C^{-(\gamma+1)} N^{\varepsilon} \quad \Rightarrow \quad C = \gamma w \tag{4.53}$$

Substituting this expression for C into (4.50) and rearranging leads to

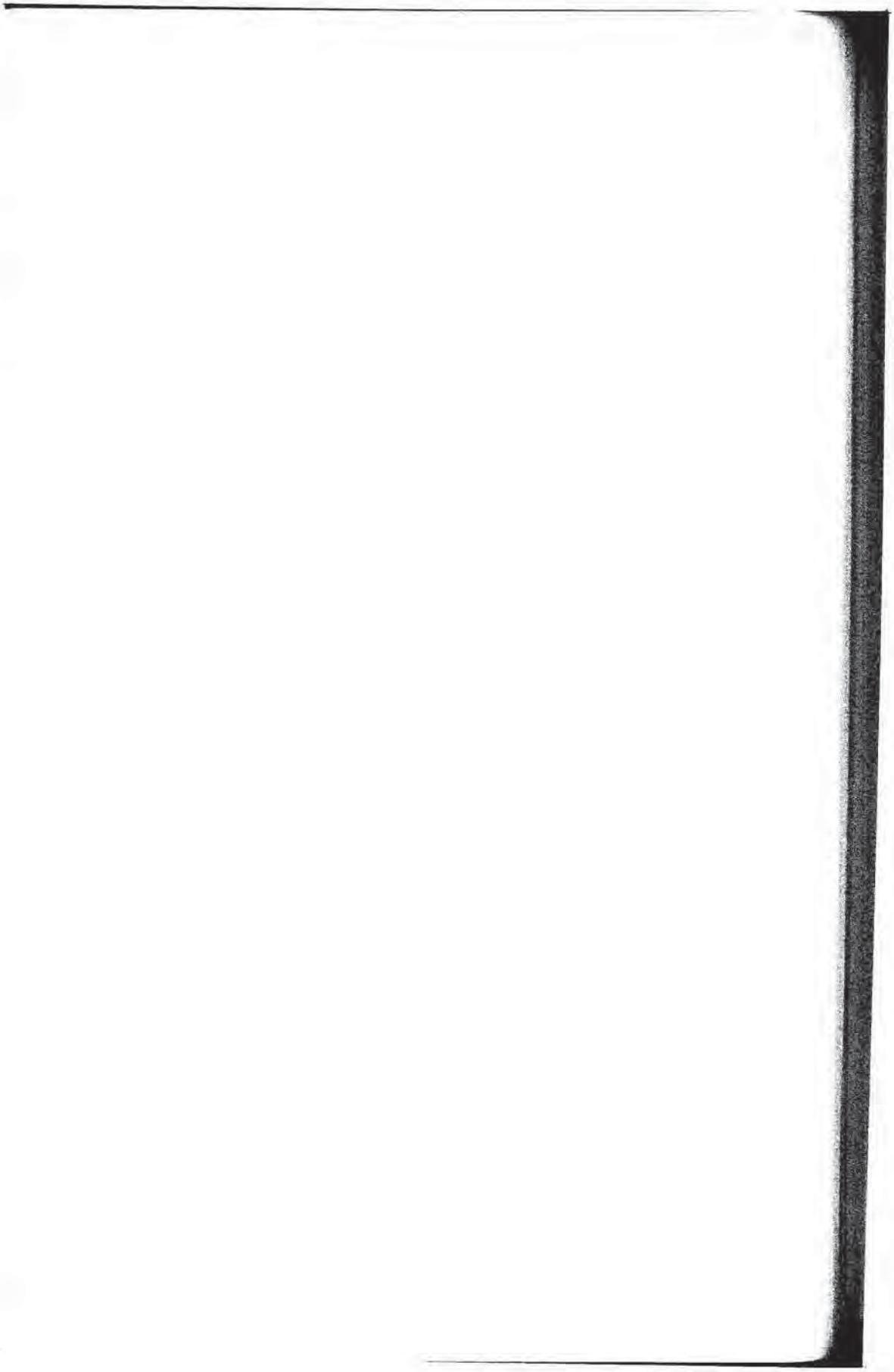
$$w = \gamma^{-\gamma/(1+\gamma)} B^{1/(1+\gamma)} N^{\varepsilon/(1+\gamma)} \tag{4.54}$$

which is (4.52). N will then be chosen to maximize $Y(N) - (1 + \gamma)wN$ where w is given in (4.54). Using the fact that (4.53) implies that $(w + C) = (1 + \gamma)w$, leads to the first-order condition of (4.51).

(4.51) says that we want to be able to estimate $\varepsilon/(1 + \gamma)$ to estimate the rate of exploitation while (4.52) says that it is exactly the parameter we would expect to estimate if we run a regression of $\log(w)$ on $\log(N)$. Of course, all the problems we have discussed earlier surrounding the estimation of (4.52) still apply: it is just that acknowledging the labor cost function causes no additional problem.

Part Two

THE STRUCTURE OF WAGES



5

The Wage Policies of Employers

To maximize profits employers would like to obtain workers at the lowest possible cost. In the models used in previous chapters, employers were constrained to set a single wage for all their workers. The choice of this wage forces the employer to trade off the number of workers (the higher the wage the easier is recruitment and retention of workers) against the profit per worker (the higher the wage the higher are labor costs). The employer ends up paying some workers more than it needs to recruit them and misses out on the recruitment and retention of other workers it could have profitably retained at a different wage. There is an incentive for employers to find alternative wage strategies to increase profits. In the jargon of economists, employers have been assumed to be a simple monopsonist, but there are incentives for them to become a discriminating monopsonist.

The ways in which employers might try to do this are the subject of this chapter. It has two main conclusions. First, that employers with market power are predicted to use wage policies similar to those observed, for example, the use of seniority wage schedules. This is confirmation of the usefulness of our approach to labor markets but it also raises the possibility that conclusions based on the simple models of previous chapters could be misleading. In the static theory of monopsony, it is well known that inefficiency is eliminated if employers can practice perfect wage discrimination although all surplus would then all go to the employer. In dynamic models of monopsony and oligopsony, matters are more complicated: chapter 3 came to the conclusion that there are situations in which increasing the employer share of the surplus may reduce efficiency; in this case wage discrimination may worsen performance. For example, if employers do manage to extract all surplus from matches with workers, then there are no incentives for workers to search or invest in human capital. The important distinction is between ex ante and ex post efficiency, a distinction that does not arise in a static model.

The chapter then argues that, in practice, evidence strongly suggests that employers are severely limited in their ability to be discriminating monopsonists. In particular, wage variation is very low among workers who do the same job: the reasons for this are not entirely clear but may well have something to do with worker demands for fairness. Although

these forces that limit wage variation are somewhat mysterious they do seem to act as a powerful constraint on the ability of employers to be discriminating monopsonists.

5.1 The Discriminating Monopsonist

In the simple Burdett–Mortensen model of section 2.4, a firm that pays all workers a single wage w is missing out on two potential sources of extra profits:

- employed workers who leave the firm when they get a better offer from elsewhere but who could be induced to stay if their current wage was raised sufficiently;
- workers employed in other firms at higher wages who could be profitably induced to move to this firm by the offer of a higher wage.

In addition, if there is heterogeneity in the reservation wages of workers (as in the modified Burdett–Mortensen model of section 3.5), then paying a single wage also misses out on profits from:

- new workers hired from non-employment whose reservation wage is below w so could have been hired more cheaply;
- non-employed workers whose reservation wage is above w but who could be profitably employed at a higher wage.

There are incentives for firms to design strategies to capture some of these untapped profits. As this means paying different wages to different workers according to their circumstances (but not their productivity which, in this chapter, is assumed identical), this is a form of wage discrimination.

Extra profits from the last two groups of workers can be captured if the reservation wage can be observed. There are obvious difficulties in doing this as it is not clear how the employer can obtain the requisite information on the reservation wage. For example, in the simple model of sections 2.4 and 3.5, the reservation wage is simply the value of leisure.[1] It is plausible to assume that the value of leisure is the private information of the worker in which case the only incentive compatible contract would be to offer the same contract to all workers.[2] To the extent that the reservation wage is correlated with some observable characteristics, we

[1] In the more complicated model of the reservation wage introduced in section 9.1, the value of leisure continues to be important.

[2] This is because employment is a 0–1 decision in the current model: if there was a continuous employment decision, for example, the choice of hours, then it is possible that, by using a wage–hours package, the employer may be able to successfully wage discriminate although this discrimination will generally be less than perfect.

would expect to see employers making different wage offers to these groups, an idea that we pursue further in chapter 7 when we discuss discrimination. But, unless the reservation wage can be perfectly predicted, the employer will be unable to capture all the surplus in this way and some workers who could be profitably employed will remain unemployed.

Now consider the strategy for dealing with workers who are already in employment. If all other firms are following single wage strategies, then it is optimal for this firm to pursue a strategy of offer-matching. When an existing worker receives a better offer from elsewhere, the employer should match that offer (as long as it does not exceed the marginal product of the worker). And, when matched with a worker currently employed elsewhere, this employer should match their wage (or pay them slightly above it) to induce them to move, again subject to the proviso that this wage offer does not exceed their marginal product.

Of course, we would expect other firms to pursue a similar strategy so let us consider briefly what would happen in general equilibrium if all firms pursued offer-matching strategies (a question analyzed by Postel-Vinay and Robin 2002). In modeling such a labor market, the most important decision is how to model what happens when an employed worker manages to match with another employer. One assumption is that the two potential employers make offers to the worker who then accepts the higher of the two. The firm that values the worker most will hire the worker at a wage equal to the productivity of the worker in the other firm. In the present model where the productivity of the worker is the same in all firms, this means that the wage will immediately get bid up to p as soon as the first external wage offer when in employment is received.[3] This may seem to be very destructive of the firm's profits as no employer then makes any further profit from the worker but the initial wage paid to workers recruited from unemployment can be adjusted downwards as workers will be very enthusiastic to get into employment. Because there is no competition among employers for workers entering jobs from non-employment, workers may do worse in a labor market with offer-matching than in the labor market with the single wage policy even though one might have thought that offer-matching would be good for workers. Workers will certainly be worse off when there is no variation in reservation wages across workers as the employers can then set the initial wage to extract all the surplus from workers.[4]

[3] Matters are much more complicated if there is heterogeneity in the productivity of firms; see Postel-Vinay and Robin (2002) for an analysis of this case. But, the conclusion that offer-matching is to the disadvantage of workers remains.

[4] We will not present a formal analysis of this labor market as it is isomorphic to the model of seniority wages that we consider later.

Although offer-matching is seen in some labor markets (perhaps most familiarly the American academic labor market), it is relatively rare. Even in labor markets that one thinks of as being highly individualistic, such as Wall Street, employers seem reluctant to engage in offer-matching: Lewis (1989: 149) describes how Salomon Brothers lost their most profitable bond trader because of their refusal to break a company policy capping the salary they would pay. Weiss (1990: 46) cites a *Wall Street Journal* article on the subject of offer-matching and suggests a number of possible reasons for its rarity.

First, he suggests that if an offer-matching strategy is pursued, workers will have an incentive to search for outside offers with the result that they leave jobs more quickly. Offer-matching is then to the disadvantage of employers. But, if the outside wage offers are treated as exogenous (so there is no bidding game for the worker among potential employers) workers end up with a wage that they would have received in the absence of offer-matching (the only difference being which employer they work for) so the incentives to search are the same both with and without offer-matching. With offer-matching, as the firm gets to keep the worker and continue to make profits from them in some situations where they would have quit in the absence of offer-matching, this argument alone cannot explain the rarity of offer-matching. But it is plausible to assume that workers also receive some non-pecuniary benefit from the job that is their private information. Then, a worker who particularly likes the job and who has no intention of moving unless they get a much better wage offer may not search much if there is no offer-matching but may search harder when there is offer-matching.

The second argument that Weiss presents is that it may be very difficult for the employer to verify outside wage offers. If, at the extreme, the employer has no ability whatsoever to observe outside offers, then any employer that instigated an offer-matching strategy would find that its workers immediately generated outside wage offers that just happened to be equal to their marginal product within the firm.[5] The appeal of this approach is that it can explain why offer-matching seems more frequent in some labor markets than others. For example, in the US academic labor market, there is an enormous amount of information about both workers and employers: a professor who wanted a raise because he/she claimed he/she had a fantastic job offer from the East Kansas Bible School of Business would lack credibility. But, offer-matching is much rarer in labor markets for unskilled workers because they are much more anonymous: a worker in Kentucky Fried Chicken who claimed they had a good outside offer from East Kansas Fried Chicken just

[5] It might be recognized that even this is not a problem if entry fees can be charged for workers to the firm: this insight is discussed in the next section.

might be telling the truth.[6] So, we might expect offer-matching to be more frequent in less anonymous labor markets where there is good information about alternative employers and workers.

Finally, Weiss suggests that offer-matching is bad for morale. The distribution of wages among the workers in the firm will reflect, not just productivity, but who has been lucky enough in exciting the interest of outside employers: it is commonly argued that such wage variation is disliked intensely by workers and is avoided by employers. We will present evidence later in this chapter that there is surprisingly little wage variation within firms which suggests that there may be some truth in this. Many economists are unhappy with assuming that workers have preferences of this sort but these economists often make the mistake of imagining that the conventional way in which the utility function is modeled with the individual caring only about what happens to themselves is anything more than an arbitrary (although convenient) assumption (for the confession of a largely neoclassical economist in this regard, see Rees 1993).

So, there are a number of powerful reasons why there may be serious limits to the offer-matching strategies that can be pursued by employers and this puts limits on their ability to act as discriminating monopsonists.

5.2 Non-Manipulable Wage Discrimination

The basic problem with the wage discrimination strategies described above is that they are extremely vulnerable to exploitation by workers when the employer's information is less than perfect. For example, paying different wages according to the worker's reservation wage or matching outside offers is unlikely to be a sensible strategy when reservation wages are unobservable, when the worker obtains an unobservable non-pecuniary benefit from the job, or when outside offers are unverifiable. The point that feasible labor contracts are severely limited by the information available to employers is one also made by Hall and Lazear (1984) who use similar arguments to the ones presented here. But, these arguments do not mean that the firm cannot practice any wage discrimination: it just means that any wage discrimination that is practiced must be based on characteristics of the worker that are non-manipulable. Chapter 7 considers discrimination on the grounds of sex. Here we consider two other characteristics of the worker that are observable: age and job tenure.

[6] In fact, to the best of my knowledge, there is no East Kansas Fried Chicken but there is a Kansas Fried Chicken along with Tennessee, Miami, Tex-Ann, Dixy, Kennedy, KCFC among others.

5.2.1 Seniority Wages

This section considers the optimal contract when the firm is allowed to pay different wages to workers according to their seniority on the job. In particular, we assume that a wage $w(t)$ is paid to a worker with job tenure t. For reasons that will become apparent, we also assume that a worker joining the firm makes a lump-sum payment of B to the employer. We assume that the employer can commit to the contract $\{w(t), B\}$ so a worker offered a contract believes it will be honored.

The important consequences of this contract structure can be conveyed in a partial equilibrium model. So, at the risk of some abuse of the notation used in previous chapters, assume that the distribution of the *value* of outside jobs is $F(V)$ and assume that this does not vary with job tenure in this firm on the grounds that workers are forced to start anew in any new job. Assume that the flow of recruits to the firm also depends on the value of the job: denote this function by $R(V)$.

The value of a job in this firm obviously depends on the tenure of the worker. Denote by $V(t)$ the value of the job to a worker with tenure t.[7] $V(t)$ must satisfy

$$\delta_r V(t) = w(t) - \delta_u[V(t) - V^u] + \lambda \int_{V(t)} [V - V(t)]dF(V) + V'(t) \quad (5.1)$$

where we have assumed from the outset that the value function is differentiable (as will turn out to be the case).

The separation rate for workers of tenure t will be $\delta + \lambda[1 - F(V(t))]$ so that the employment of workers of tenure t, $N(t)$ must satisfy the following differential equation:

$$N'(t) = -[\delta + \lambda[1 - F(V(t))]]N(t) \quad (5.2)$$

Employment of new workers, $N(0)$ must be given by the new recruits so that

$$N(0) = R(V(0) - B) \quad (5.3)$$

which reflects the fact that the entrance fee has not been included in the specification of the value function in (5.1). In a steady state, profits, Π, will be given by

$$\Pi = \int_0^\infty [p - w(t)]N(t)dt + BN(0) \quad (5.4)$$

The employer wants to choose $\{w(t), B\}$ to maximize (5.4) subject to (5.1)–(5.3). The solution to this problem is contained in the following proposition.

[7] Note that the argument of the value function used here is job tenure, not wages as in most other parts of the book.

Proposition 5.1. *An optimal wage strategy for the firm is to set the wage equal to marginal product, p, at all tenures and choose the "entry fee" B to maximize*

$$\Pi = BR(V^* - B) \tag{5.5}$$

where V^* is the value of the job to the workers when the wage is always equal to the marginal product.

Proof. See Appendix 5.

There is a very simple intuition behind this result. In the situation where the employer is a simple monopsonist setting a single wage for all workers with no entry fee, the level of the wage is determined by two conflicting pressures. On the one hand, high wages reduce the profit per worker but, on the other hand, high wages reduce quits. As soon as seniority wage schedules and entrance fees are allowed, high wages can be used to deter quits without necessarily reducing profits because the benefit to the worker of higher future wages can be captured by the firm in the form of higher entrance fees.

It should be noted that the optimal contract is not unique. For example, if a wage below p is high enough to prevent all quitting to other employment, then that wage can also be optimal together with a reduced value of B.[8] Stevens (1998) provides a more thorough analysis of the set of optimal wage policies and shows that all optimal policies must have the same turnover outcomes and the same value of a job for workers joining the firm.

The idea that the wage structure can be used to deter quits is not new. Ioannides and Pissarides (1985) present a two-period model of a monopsonistic labor market, where it is optimal for older workers to be paid their marginal product and younger workers to be paid less. They argue that their result shows that it is optimal for wages to increase with job tenure. However, Proposition 5.1 shows that the optimal wage structure in a model with more than two periods does not really resemble a smooth relationship between wages and job tenure. It is optimal to pay all workers a wage equal to their marginal product but to make new workers "buy" their jobs: this point is obscured in a two-period model.

Proposition 5.1 shows that the employer can increase profits by moving from a contract in which the only payment between employers and workers is a constant wage to one in which this constant wage is supplemented by an entry fee. But, one might also wonder if the firm cannot do even

[8] The optimal contract in this case can also be thought of as the optimal contract in the presence of offer-matching where B is interpreted as the value of the low initial wage paid to workers recruited from unemployment.

better by introducing a richer form of labor contract. For example, one might wonder whether an improvement could be made if lump-sum payments were made between worker and firm not only when the worker joins the firm but also when the worker leaves the firm, that is, to consider severance payments and/or bonding arrangements. To investigate this, let us assume that the wage contract can be represented by $\{w(t), S(t), B\}$ where $S(t)$ is a payment made by the worker to the firm whenever they leave it (which could be negative). One might want to allow for these severance payments to differ by the reason why the worker leaves the firm. But it seems reasonable to assume that this is not feasible as a worker could always artificially induce a period of unemployment before starting a new job if it was beneficial to do so. The following proposition shows that allowing for severance payments introduces no new freedom for the employer.

Proposition 5.2. *An employment contract with severance payments is equivalent to one without. In particular, the contract $\{w(t), S(t), B\}$ with severance payments is exactly equivalent to the following employment contract without severance payments:*

$$\tilde{w}(t) = w(t) - S'(t) \tag{5.6}$$

$$\tilde{B} = B + S(0) \tag{5.7}$$

Proof. See Appendix 5.

The intuition for this result (which can also be found in Stevens 1998) is very simple. First consider the case where the severance payment is constant. A worker joining the firm knows that he/she will leave it at some point, thus incurring the severance payment, so that the real effective entrance fee is the actual entrance fee plus the severance payment.[9] If the severance payment is not constant, then one can think of the reward to staying with the firm for an extra period as being the wage payment minus the change in the severance payment which is (5.6). The effective entrance fee is now the actual entrance fee plus the initial severance payment.

The optimal contract derived here is not really close to anything that is observed in most of the labor market. Seniority wage schedules do exist, but they do not have the extreme form suggested by the model derived

[9] This uses the assumption that agents do not discount the future. One might reasonably wonder if this is important: it is not if both workers and employers have the same discount rate although the equations need modification. But, equivalence will fail if the discount rates of worker and employer differ.

above. In particular, entrance fees are extremely rare. To understand the reasons for this, we can draw here on discussion of another labor market model, the shirking version of the efficiency wage model (Shapiro and Stiglitz 1984) which attempts to explain the existence of involuntary unemployment. This debate is about whether firms can get workers to post bonds that are forfeited in the case of bad performance (Carmichael 1985).

One argument about the limit to entrance fees is that the access of workers to capital and insurance markets is often limited and workers want to smooth consumption both over time and across states of nature (although we have implicitly assumed they only care about the expected discounted income stream). So, workers might simply not be able to afford the entrance fee demanded by the first-best contract.[10] In addition, some of these workers might lose their jobs very quickly so will not have had time to recoup any of the gains. To some extent one can avoid these problems by judicious use of severance payments. For example, one could implement the first-best contract without any entrance fee by setting a severance payment equal to the optimal entrance fee and then paying workers their marginal product. Workers are not then required to put any money up front into the firm. But, they are still exposed to substantial risk as a worker who is unlucky enough to be forced to leave the job very quickly has a large liability to the employer and their ability to pay will again be limited by capital market imperfections.[11] However, large severance payments are problematic in many countries where anti-slavery laws make courts reluctant to enforce any labor contract that looks like it gives the worker no choice but to remain with their employer.

Another problem with the use of entrance fees is the potential for abuse of the system by employers. There are obvious incentives for the employer to take the entrance fee from the worker and then fire them. This potential problem has been addressed in the context of the shirking model by Macleod and Malcomson (1989) who argue that this is not a serious problem. Put in the context of the current model, their argument is essentially that employers do not have a positive incentive to fire workers after they have paid their entrance fee as workers are being paid less than their marginal product and reputation effects will prevent employers from behaving in this way. But, their model is too optimistic in one important way. It is implicitly assumed that firms have sunk some investment in the

[10] Burdett and Coles (2001) analyze a model of the type considered here with risk-averse workers.

[11] Better insurance could be achieved if the severance payment could differ according to whether the worker leaves for unemployment or another job as the move to the new job, being voluntary, must make the worker better off whereas the involuntary move to unemployment does not.

creation of the job so that all employers have something to lose if they get a reputation for cheating workers. The assumption that creating a job is costly is a reasonable assumption when the employment contract is a constant wage as workers know that any employer offering this presumably has some productive job behind it or else they would not make any profit. But, imagine what would happen if the form of employment contracts was such that workers were asked to put money up front into a job. There is then an incentive for rogue employers to try to recruit workers even if they have no real job. As the set-up costs for a fake job are likely to be very low, the market would become swamped with fraudulent employers who promise workers jobs, take their entrance fees, and run. So, it does seem a reasonable argument that any employer that tried to offer an employment contract with large entrance fees would be treated with the utmost suspicion by workers. In fact, many newspapers, London lamp posts (and increasingly e-mails) contain advertisements for jobs which promise very large rewards—so large, that I would be well advised to apply for them. On investigation (something I would recommend just for the curiosity value) these "jobs" typically require the purchase of materials or training information (which are equivalent to entrance fees as I suspect they are substantially above the marginal cost of these materials) and, if one is unwise enough to pursue this further, one often finds it is very hard to get any money at all out of the "employer" as one's work is often deemed unsatisfactory. I suspect that I and most other workers are wise in their avoidance of these tempting job offers.

One simple way of modeling the constraints imposed by capital market constraints and employer opportunism is to impose a restriction on the optimal contract that the worker can never be in debt to the firm, that is,

$$-B + \int_0^t w(s)ds \geq 0 \qquad \text{for all } t \tag{5.8}$$

(5.8) requires the total cumulative payment from firm to worker at any tenure t to be non-negative. This requires that $B \leq 0$ and that $w(t)$ must start off positive. One optimal contract with this restriction is given by the following proposition.

Proposition 5.3. *If workers can never be in debt to employers (i.e., the constraint (5.8) is imposed on the set of feasible contracts), then it is optimal for the employer to offer a zero wage for a certain period of time and then to raise the wage to the marginal product.*

Proof. See Appendix 5.

This type of contract (which Stevens calls a step contract) begins to bear some resemblance to what is commonly observed where workers are offered a low starting wage for a probationary period and the wage is then raised to their marginal product.[12] Again, Stevens (1998) has shown that the optimal contract is not unique but that employers can never do better than using a step contract.

A final problem with the use of entrance fees is that a lot of information is needed to set them. In particular, one needs to know the value of the job to workers when they are paid their marginal product which needs knowledge of their quit rates. There is a certain amount of empirical evidence (Weiss 1984; Farber 1994) that there is important individual heterogeneity in quit propensities. An employer's information about individual quit propensities is likely to be very imperfect so that it is reasonable to think of the quit rate of workers as being their private information in which case the entrance fee cannot be made contingent on it. A contract with a high entrance fee now has the rather undesirable feature that it is unattractive to workers with high quit propensities.[13] Analyzing the optimal contract in this case is rather difficult. So we will proceed by considering a simple example.

Suppose that, once in employment, workers only leave the job to exit the labor force and there is no potential for job-to-job mobility. A worker with quit rate δ in employment then has an expected revenue of (p/δ). If the supply of workers with quit rate δ is given by $R(V, \delta)$ then profits from δ workers will be given by

$$\Pi = \left(\frac{p}{\delta} - V\right)R(V, \delta) \tag{5.9}$$

Suppose that δ is observable. Then V will be chosen to maximize (5.9): denote the choice by $V^*(\delta)$. The optimal contract is not uniquely determined: any $\{w(t), B\}$ which satisfies

$$\int_0^\infty w(t)\exp(-\delta t)dt - B = V^*(\delta) \tag{5.10}$$

will do the job. In our previous model, this indeterminacy was partially resolved by the fact that the employer also wants to deter quits.

Now suppose that δ is not observable and that only a single contract $\{w(t), B\}$ can be offered by the employer.[14] The employer can obtain the

[12] One could also impose a subsistence requirement that workers wages cannot fall below a strictly positive amount. This would have straightforward consequences for the optimal contract.

[13] This disadvantage may be a blessing in disguise if there are training costs. Salop and Salop (1976) present such a model in which entrance fees are used to discourage workers with high quit propensities.

[14] We are ignoring here the possibility of self-selecting contracts being offered by the firm.

first-best if it can find a contract that satisfies (5.10) for all δ. This is possible in some cases. For example, suppose that $R(V, \delta) = r(\delta)V^{\alpha}$ so that the supply is isoelastic with the same elasticity for all δ. Then it is simple to show that $V^*(\delta) = [\alpha/(1 + \alpha)](p/\delta)$. This is attainable even if δ is not observable if a constant wage is offered with $w = [\alpha p/(1 + \alpha)]$ and no entrance fee. Of course, with other specifications of the supply curve a different result would be obtained, but the isoelastic specification is the most natural place to start. So, in the absence of competition for workers from other employers, it is likely that the employer will offer a rather flat wage–tenure schedule. With competition from other employers, it is likely that the employer will try to choose the wage structure to hit two conflicting targets: a steep wage–tenure profile to deter quits but a flat one so that workers with a high quit propensity are still attracted to the firm.

The optimal contract in this case is likely to be a mess but that is probably the right conclusion. The important point is that the degrees of freedom available to the employer in designing its wage policy are unlikely to be sufficient to meet all its conflicting objectives. Different employers are likely to find different wage policies optimal and that accounts for the very considerable diversity that is observed.

It is also worth mentioning the ingenious discussion of Stevens (1998) about the general equilibrium features of a monopsonistic labor market in which employers can use these policies. She shows that the Burdett–Mortensen conclusion that there must be wage heterogeneity across firms disappears and all firms offer equivalent wage–tenure contracts. However, wage dispersion remains as identical workers receive different wages within the firm according to their seniority. However, Burdett and Coles (2001) show that employer heterogeneity in equilibrium contracts is re-established if workers are risk averse.

5.2.2 Wage Discrimination by Age

There are also incentives in the Burdett–Mortensen model for employers to pay different wages to workers of different ages. The reason for this is that the non-employment rate of workers is declining with age so that an older worker is more likely to be recruited from other firms and to be in higher-wage jobs. Both factors encourage the employer to pay higher wages to older workers. But, as long as some workers of every age are recruited from non-employment it will never be optimal for the employer to pay a worker of any age their marginal product.

In practice, the non-employment rate by age is fairly flat after the age of 25 so that this effect might not be expected to be very important for adults. But for younger workers, it is potentially important and could account for the widespread practice of paying distinct (and lower) youth wages.

5.3 Empirical Evidence

In this section, we present empirical evidence in support of two of the main claims made above. First, that employers do have wage structures that are broadly consistent with the model, notably that they do tend to pay higher wages to older and more senior workers. Secondly, that the wage policies pursued by employers show evidence that they are not flexible enough to attain all objectives so that they fall some way short of attaining their ideal of being perfectly discriminating monopsonists.

5.3.1 Wages, Age and Tenure Within the Firm

It is well known that there are generally positive correlations between wages, age, and tenure in the labor market as a whole. The next chapter examines the hypothesis that part of the cross-sectional returns can be explained by the fact that, in a labor market with frictions, it takes time to find a good job and, once one has found it, one tends to keep it. But, that is not what interests us here. We are interested in whether, *within firms*, employers pay higher wages to older or more senior workers.

There are obviously other theories as to why older and more senior workers get paid more. The most popular explanation is based on human capital theory: that older workers are paid more because they are more productive, and that more senior workers have accumulated more firm-specific human capital and it is optimal for the employer to give workers a share of these rents.

Although human capital theory is the accepted way of thinking about the returns to tenure, the reasoning behind it is not as sound as one is often led to believe. Strictly speaking, an employer has no need to pay a worker in a competitive labor market a wage any higher than what they could obtain in the open labor market, that is, "if all training were completely specific, the wage that an employee could get elsewhere would be independent of the amount of training received. One might plausibly argue, then that the wage paid by firms would also be independent of training" (Becker 1993: 41–42). But Becker (1993) argues that if the worker was only paid his/her outside option then, if he/she quits, the rents lost by the employer would not be taken into account. But, in the perfectly competitive world of full information where outside options are perfectly known and costlessly available, this argument is unpersuasive. If the worker's outside option rises and he/she threatens to quit, then the simple thing to do is to match the outside option. This is not only simple, but efficient. There is no reason why the wage paid should be determined

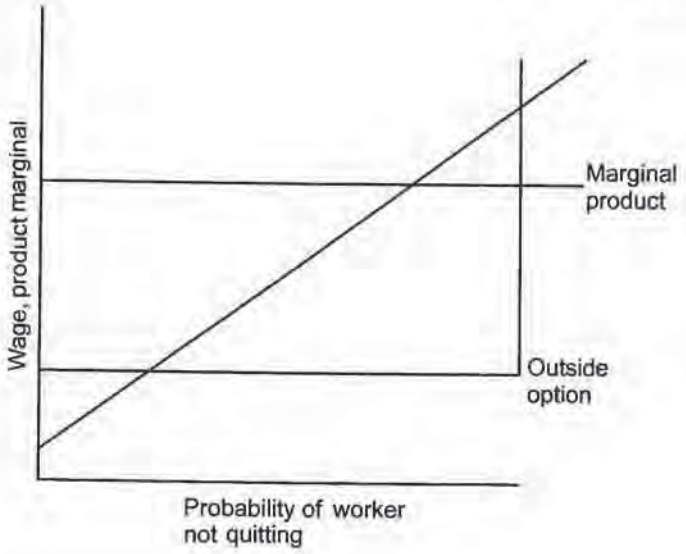


Figure 5.1 The Hashimoto model.

by anything other than the outside option, that is, we would not expect it to be related to the extent of job-specific human capital. It is simple to understand this in terms of figure 5.1. The quit probability of the worker is 1 for a wage below the outside option and zero for a wage above it leading to the "backward-L" shaped labor supply curve of figure 5.1. The profit-maximizing thing to do is then to pay the outside option and this is independent of the productivity of the worker in this particular firm. One way to understand this is to think of the retention function as being the labor supply curve facing the firm and to recognize that this labor supply curve is the competitive one as it is perfectly elastic at the outside wage.

Becker's informal argument about the sharing of the returns to specific human capital is normally argued to be supported by the more formal model of Hashimoto (1981). The optimal contract in his model is derived by assuming that the outside option is either unobserved by the firm or that it is simply too costly to write a contract contingent on it. The retention probability of the worker is no longer the right angle of figure 5.1 but a smooth increasing function of the wage as also drawn in figure 5.1. Hashimoto then shows that the optimal wage to pay is increasing in the productivity of the worker.[15]

But, it is important to realize that this result is only obtained by effec-

[15] His model is actually more complicated as he has uncertainty about the returns to specific human capital within the firm so that he can also focus on the possibility of firing.

tively assuming that the labor market is monopsonistic, that is, that the supply curve of labor facing the firm is not perfectly elastic as the higher the wage paid, the higher the probability of retaining the worker, and, hence, the higher the level of expected employment. Becker (1993: 44) admits as much: "the likelihood of a quit is not fixed but depends on wages." But, once one allows this, all other sorts of problems emerge. For example, one could argue that the labor market remains competitive because, at the moment the worker is hired, there is a well-defined outside option which is the market price for the worker. But, if one wants to assume that the outside option of the worker is uncertain ex post, then it seems natural to argue that it is also uncertain at the point of hiring and then labor markets are truly monopsonistic with all that that implies. For example, there seems no good reason why the outside option available to the worker in the Hashimoto model should be his /her marginal product in other firms, yet that is the assumption made. Not for the first time, we find monopsony arguments being used when convenient but a failure to fully think through the implications of the assumption being made. It is much better to think of the Hashimoto model as showing why, in a monopsonistic market, firms in which workers (for whatever reason) have higher productivity will pay higher wages; we return to this subject in chapter 8.[16]

However, there is some evidence that we should be skeptical of the statement that more experienced, senior workers are indeed more productive. To test this hypothesis one obviously needs data on the productivity of workers independent of the wage and that is hard to come by. For example, Medoff and Abraham (1980, 1981) found that in two US corporations the correlation between the pay of professional and managerial workers and their experience was much stronger than the correlation between performance and experience.

There are other non-competitive explanations of why wages may vary with age and tenure independent of productivity. For example, shirking versions of efficiency wage models have suggested that an upward-sloping wage profile is a cost-effective way of providing incentives for workers (Lazear 1979; Akerlof and Katz 1989). There is no way that the available data can sort out this hypothesis from ours as they are extremely similar. In both cases the wage policy of the firm is determined by the struggle between two competing problems: the desire to lower labor costs and the desire to ease recruitment and retention in our case, and the desire to provide incentives not to shirk in the incentive models. In fact, we freely

[16] One implication of this discussion is that even if one did show, as Brown (1989) claims, that firm-specific wage growth occurs almost exclusively during periods of on-the-job training, this could legitimately be taken as evidence of the existence of monopsony rather than the validity of a competitive human capital story as one would expect employers to raise wages when productivity is raised by training.

used ideas from this efficiency wage literature in our discussion of the constraints on the wage policy of the firm.

5.3.2 Constraints on Wage Policies

The evidence of the previous section suggests that the variation in wages within firms cannot all be explained by variations in productivity as the traditional human capital approach would suggest. This leaves open the possibility that employers do practice some wage discrimination and are not simple monopsonists. The natural next question is whether employers manage to get sufficient flexibility in their wage structures to be usefully seen as perfectly discriminating monopsonists. To address this question directly would require information on the productivity of workers and their alternative sources of employment in the labor market and this is simply not available. So, this section takes another approach to the question.

There is considerable evidence that wage policies within firms are not individualized to any great degree: there seem to be powerful forces making the terms and conditions offered to workers in a given job within a firm very similar. Given that it is plausible that there is considerable heterogeneity among the workers within a firm, this evidence implies that the firm is not able to take account of the specific circumstances of an individual worker in determining his/her contract of employment and, hence, is not able to get close to being a discriminating monopsonist. This view of the lack of individuality in employment contracts is not new: Webb and Webb (1897: 281) wrote that "the most autocratic and unfettered employer spontaneously adopts Standard Rates for classes of workmen, just as the large shopkeeper fixes his prices, not according to the higgling capacity of particular customers, but by a definite percentage on cost."

We present two pieces of evidence on the lack of individuality in employment contracts. First, we have data from a survey of wages in vacancies posted in UK job centers conducted by the Manchester Low Pay Unit in the spring of 1994. On 1 September 1993 the Wages Councils that had previously set minimum wages in a number of low-paying sectors were abolished and (except in agriculture) the United Kingdom had no minimum wage. Because of the standard provisions of employment law, employers were unable to cut wages for existing workers. But, there were no such restrictions placed on the wages that could be offered to new recruits. So, when the Manchester Low Pay Unit conducted a survey of wages offered to new recruits in sectors that had been covered by the Wages Councils there was no reason why the minimum wage that had previously been in force would have any particular salience. Yet, what they found was a spike of wage offers at or very close to the minimum wage. The data are shown in table 5.1. On average, 18% of vacan-

TABLE 5.1
Starting Wages in Old Wages Council Industries

Wages Council	*Hotels/ Restaurants*	*Cafes*	*Pubs/ Clubs*	*Food Retail*	*Non-food Retail*	*Clothing*	*Hair-dressing*	*All*
Minimum wage at abolition	2.92	2.99	5.01	5.17	5.15	2.71	2.88	n.a.
Number of vacancies	702	182	469	293	471	270	127	2514
Average wage	5.16	5.07	5.10	5.19	5.32	5.01	5.16	5.16
Percent at old minimum	15	5	45	15	10	3	12	18
Percent at closest "focal" wage	1	27	12	4	10	6	2	8
Percent within 1% of old minimum	16	35	59	33	19	8	16	27
Percent within 5% of old minimum	56	59	81	55	53	23	42	56

Notes.
1. Source is data collected in spring 1994 by Manchester Low Pay Unit (Cox 1995). The data are the wages at which vacancies are advertised in Manchester job centers.

cies were at exactly the old minimum wage although the percentage varies considerably from 3% in clothing manufacturing to 45% in pubs and clubs. It should be noted that the minimum wages were set at hourly wages which were not round numbers at which one typically sees spikes in wage distributions. For example, the minimum wage was set at £2.99 in cafes and 5% of vacancies offered exactly this, but one never normally sees hourly wages reported that end in "99" (e.g., in the US CPS only one in a thousand workers report an hourly wage ending in a "99"). But unsurprisingly there was a larger spike at £3.00 per hour so the next row reports the size of the spike at the nearest "focal" wage, that is, an hourly wage ending in a zero or a five. The last two rows report the fraction of wage offers within 1% and 5% of the old minimum: these are large numbers. Even in the industries where there is a small spike close to the old minimum, this is because there is a very large spike at some other wage (25% at £3 per hour in cafes). All this suggests a picture in which initial wages are not very sensitive to the characteristics of the individual job applicant.

But, this is about the wages posted by employers at job centers: it says nothing about wages actually paid. But we do see a similar picture among wages actually paid. Machin and Manning (2002) report the results of a survey of workers in residential homes for the elderly on the south coast of England in 1992 and 1993. There was no minimum wage in this sector, nor any unions or prospect of unions, so the wage structures we observe are the "free" choices of the employers. Yet, we still see a lack of individuality in the wages paid. 26% of workers work in the 31% of firms that have only a single hourly wage paid to all their care assistants (the main occupation). And the fraction of the total variation in wages that is within-firm is, as table 5.2 reports, much lower than the fraction of any other variable on which we have data (age, hours, and job tenure).[17] This remains true even if we have very detailed geographical controls and if we restrict attention to the larger firms in the sample.

All this adds up to a picture of a labor market in which there are serious constraints on the ability of firms to individualize wages and that these constraints will act as a limitation on the ability of firms to act as discriminating monopsonists. The obvious next question is "why don't firms individualize wages?" There are a number of possible explanations for this. Webb and Webb (1897) saw "practical convenience and the growth of large establishments" as the reason, while a number of authors (e.g., Akerlof and Yellen 1990; Solow 1990; Bewley 1999) have suggested that morale of workers suffers if the wage policy is felt to be unjust.

[17] It is important to remember that we are looking at wage dispersion in a very tightly defined occupation. The wage gap between managers and cleaners is almost certainly largely within-firm.

TABLE 5.2
Wage Dispersion Within Firms: Evidence from UK Residential Nursing Homes

		Log Wage	*Log Age*	*Log Tenure*	*Log Hours*
All workers					
No controls	1992	0.74	0.25	0.32	0.36
	1993	0.80	0.20	0.30	0.36
Area controls	1992	0.68	0.24	0.33	0.35
	1993	0.76	0.21	0.29	0.35
Town controls	1992	0.57	0.18	0.26	0.24
	1993	0.65	0.15	0.21	0.23
Workers in firms with more than five workers					
No controls	1992	0.72	0.20	0.28	0.31
	1993	0.80	0.16	0.27	0.29
Area controls	1992	0.65	0.19	0.30	0.30
	1993	0.75	0.17	0.25	0.28
Town controls	1992	0.48	0.11	0.20	0.16
	1993	0.59	0.10	0.15	0.12

Notes.
1. The data in this table come from a survey of workers in residential care homes for the elderly on the south coast of England conducted in 1992 and 1995. The results presented here refer only to those workers classified as "day care assistant." See Machin et al. (1993) and Machin and Manning (2002) for more details of this data set.
2. The number in each column represents the fraction of total dispersion in the relevant variable (log wage, age, tenure, or hours) that is between-firm.

Although the recent literature has primarily emphasized concerns for fairness among workers, concerns for fairness among employers may also be important. Some older work (e.g., Reynolds 1951) argued that employers see other employers who actively poach workers as being engaged in unfair competition to some extent and this could also act to preserve company wage policies.[18] This requires at least some implicit collusion among employers, and we lack any study of the extent of such collusion in modern labor markets, although most economists probably think it rather limited. And, the evidence of Freeman and Medoff (1984) that intra-firm wage dispersion is reduced by the activities of unions suggests that the preferences of workers is important.

Another possible problem is that once wages become individualistic, workers may also realize they have some bargaining power because of

[18] Consider the following quotes from Reynolds (1951): "each personnel manager knows that, if he steals a worker today, someone else will steal from him tomorrow, and all have an interest in playing by the game" (p. 51), and "the more significant meaning of competition is impersonal rivalry in which each employer establishes terms of employment designed to attract the number and types of workers he wants" (p. 216).

labor market frictions and use this to obtain higher wages for themselves. This is the assumption usually made in the matching literature (e.g., Diamond 1981, 1982; Pissarides 1985). Peters (1991) has shown that firms may be better off if they could post wages in advance; the problem being whether this is credible. Non-individualistic wage policies could be one way of sustaining credibility. Firms want to get a reputation for not responding to worker attempts to use their bargaining power to get higher wages. Because an employer typically has more than one worker, there is an incentive to build such a reputation while in many labor markets where workers are essentially anonymous, there is little incentive for workers to build a reputation for not accepting employer-dictated wage policies. According to this line of argument, non-individualistic wage policies should be seen as a way for employers to ensure that they retain control over wage-setting.[19]

The non-individualistic nature of employment contracts has other important implications that are not pursued here. For, if the wages of new recruits are tied to those of existing workers and the wages of existing workers cannot be reduced, this could be a powerful source of nominal and cyclical wage rigidity forcing the burden of adjustment to variations in demand onto employment rather than wages. Given this evidence one might think of constructing models of the labor market in which the wage offers of employers are not altered with the productivity of workers. Manning (1994b) pursues this idea further but we will not follow that line here.

5.4 Conclusions

In this chapter we have argued that employers do have incentives to choose wage policies in an attempt to become discriminating monopsonists. We see evidence of these pressures in the observed wage policies of firms, notably increased wages for senior and older workers. But, we also see evidence of severe limitations to the ability of the employer to be a perfectly discriminating monopsonist. Private information of workers (about reservation wages, non-pecuniary benefits from jobs, and outside offers) makes it difficult for employers to design wage contracts that extract all the surplus from workers. In addition, there seem to be powerful forces limiting wage variation across workers within jobs. Ironically it may be these forces that make the model of the simple monopsonist setting a single wage not such a bad approximation for thinking about

[19] Ellingsen and Rosen (2002) analyze a model in which some firms decide to post wages and others to negotiate wages with their workers, both forms co-existing in equilibrium.